The Circle of Life for LLMs. Was the Reaction to DeepSeek Justified?

Stephane H Maes[1]

February 5, 2025[2]

Abstract

Since the release of DeepSeek Large Language Models (LLMs) and free desktop and mobile apps, the industry, the investors, and the media have reacted with alarm, surprised that a Chinese startup—despite operating on a low budget and with limited access to specialized AI hardware—could surpass the latest ChatGPT models with reasoning capabilities. This has led to geopolitical concerns about threats to U.S. technological dominance, and the effectiveness of AI chip sanctions imposed by the U.S. on China. Investor confidence in leading U.S. tech companies involved in AI, AI hardware, and AI/cloud hosting has been shaken, contributing to a significant stock market drop on January 27, 2025.

In this paper, we argue that while the success of DeepSeek V3 and R1 is remarkable, it does not signal the decline of any major player. Instead, it is a natural progression of how LLMs and generative AI function. Most LLM providers, of a same LLM generation, rely on similar algorithms, big-data pools, and development techniques, meaning that models tend to converge in performance once their methodologies become public. Whether using proprietary or open source foundations, different starting points often lead to LLMs of comparable capabilities for a same generation. Techniques such as model distillation and reinforcement learning further enable the reduction of model size, data requirements, and hardware constraints. As a result, each time a model is developed, it can be replicated, closely matched, or even surpassed soon after—sometimes with significantly lower effort than the original, or with a significantly smaller set of parameters. This cycle of life will continue as long as LLMs remain a competitive field, by opposition to a commodity, and until new AI approaches beyond generative AI emerge, or the old AI reemerges.

We anticipate such a pattern to continue: new models will be matched and overtaken by (nimbler) competitors, while major providers respond with the next iteration of improvements—repeating the cycle. Open source models, in particular, have the advantage of drawing from broader communities and collective innovation, making it increasingly difficult for proprietary models to maintain a lasting edge. As development costs rise, it will be interesting to see whether proprietary models can sustain their dominance or whether they, too, will need to integrate open source strategies.

Ultimately, there is, and was, no reason for panic or hasty divestment. AI may be in a bubble, but if it bursts, it will not be because DeepSeek outperforms OpenAI's latest model. Instead, the real challenges facing LLMs and GenAI lie elsewhere. The path to AGI is likely beyond current LLM-based approaches. While AI agents may extend the viability of generative models for some time, factors such as the finite availability of high-quality digitized training data and the risks of model collapse due to synthetic data contamination pose more significant long-term threats. That said, if LLMs are not the future of AI, there is little reason to be concerned about new players mastering them.

———

[1] shmaes.physics@gmail.com

[2] Early analysis draft: on January 27, 2025.

Cite as: Stephane H. Maes, (2025), "The Circle of Life for LLMs. Was the Reaction to DeepSeek Justified?", https://zenodo.org/doi/10.5281/zenodo.14838733, https://shmaes.wordpress.com/2025/02/15/the-circle-of-life-for-llms-was-the-reaction-to-deepseek-justified/, February 5, 2025.

1. Introduction

1.1 The Rise of DeepSeek: A Case Study

1.1.1 What happened?

DeepSeek, is a Chinese AI startup, which has rapidly become a major player almost overnight in the Generative AI (GenAI) landscape, challenging established giants like OpenAI and other dominant GenAI LLM providers [1]. Founded in June 2023 by Liang Wenfeng, DeepSeek took the tech world by storm with its innovative and cost-effective approach to Large Language Models (LLMs).

In December 2024, DeepSeek-V3 model was released [2], showcasing significant improvements in performances and efficiency over its past DeepSeek V1 (released on November 29, 2023 [1]) and V2 models (Released in May 2024 [3]). It was released as mobile apps on January 10, 2025 [2]. Then, on January 20, 2025, just a day before Stargate Joint Venture Project was announced [4], DeepSeek publicly released the DeepSeek-R1 model, a 671-billion-parameters open source reasoning AI model. This model boasts performances comparable to OpenAI's GPT-4o, and o1, but is trained at a significantly lower cost (~ US $6 million vs. US $100+ million for GPT-4) and requires less computing power [2].

Stargate LLC is a $500 billion investment in AI infrastructure by OpenAI, Oracle, SoftBank, and MGX [4], announced at a White House press conference on January 21, 2025.

On January 27, 2025, DeepSeek-R1 app surpassed ChatGPT, as the most downloaded free app on the iOS App Store in the United States [2]. The news of this achievement triggered an 18% drop in Nvidia's share price on the same day, raising concerns about the impact of DeepSeek's efficiency on the demand for the powerful, high-end AI chips [2]. The news also rattled Wall Street, with Nvidia losing nearly $600 billion in market value. This massive market reaction, wiping out nearly a trillion dollars in market value from Nvidia and other major US technology companies, highlights the potential disruption DeepSeek poses to the existing AI landscape and all the companies invested in it [1].

As the icing on the AI cake, DeepSeek also released Janus Pro as image generator on January 31, 2025, matching or bettering Dall-E 3 [8,9].

So, DeepSeek's cost-effective approach challenges the necessity of massive investments in AI infrastructure as proposed by the Stargate Project [4], and the economic soundness of LLM AI providers like OpenAI, Anthropic, Google and Microsoft, as well as part of the growth prospects for cloud providers like Microsoft, Google, Amazon, IBM, and Oracle. .

1.1.2 Investors and Technologists concerns

Investors and technologists, anxious over the idea, promoted in particular by OpenAI, that the LLM models, that GenAI requires more and more money and power to cover training and execution cost [5,7]. There is now an (at least) apparent proof that it is possible to train and perform better or at least on par with the best US LLM models,

at a fraction of the cost, faster, and without the ever growing battery of AI specialized hardware, and computing resources.

So, is this a wakeup call for the competitiveness of US tech. businesses like OpenAI, Nvidia and Cloud providers? Does it question the soundness of capital intensive initiatives like the Stargate Project?

OpenAI, especially thrived on hype or the need for more and more investment to bring us better LLMs and allegedly give us PhD level GenAI [32], then presumably soon, really soon, AGI[3] [5]. Could it be that none of that is needed and all this money has/would have been wasted? Could many other startups and enterprises, follow suit using the same or similar paths to DeepSeek, and develop their own models?

Open sourcing the large language models[4] also means opening the floodgate to many more players able to similarly catch-up or beat the current incumbents' business models.

Also, Nvidia has thrived, pivoting with great luck from Gaming hardware, on selling high-margin GPUs essential for AI training. It may appear to face a serious threat. If DeepSeek's low-cost model gains widespread adoption, the demand for ultra-high-end hardware could decline, challenging Nvidia's dominant market position and potentially impacting its $3 trillion market cap. Think fine tuning and training on low cost RISC-V chips of commercial GPUs.

Cloud Infrastructure providers have also thrived on providing AI-powered tools and APIs and hosting LLMs, with computing, storage and AI specialized hardware, for training and deploying. If the hardware requirements, and training as well as operational/execution computing requirements drop, the demand of AI in terms of cloud chips and computing resources also drops. That could be challenging for tech players like Microsoft, Google, Amazon and Oracle, since a lot of their growth projections and revenue forecasts relies on their estimation of the enterprise customers AI needs.

However, while the global technical advances remain fluid, we do not agree that panic over US loss of AI global leadership is the realistic and pragmatic understanding of the nuanced competitive landscape.

1.1.3 Geopolitical considerations

Geopolitical concerns exacerbate existing AI challenges.. It seems that that US sanctions on advanced chips in general, and AI chips in particular did not work that well, and may be useless. Indeed DeepSeek allegedly trained its LLMs on older generation Nvidia AI chips [7,41,52,55]: it may not need the sanctioned technologies, or can access equivalent ones in China [5], granted that the ones they used were older generation, possibly acquired pre-sanctions, and that the use of only small amount of chips probably result from the pressure of the sanctions; a mixed bag conclusion on the sanction effectiveness, and an illustration of the possible unintended consequences of the practice of sanctions.

The superiority of US technology, especially in AI, seems to also have taken a credibility hit. DeepSeek algorithms and widely used methodologies seem to have beaten the best AI companies. China seems to have caught up, something already obvious if we consider the volumes of AI graduate, including in US, as well as patents and publications from China.

[3] It is not going to happen with GenAI and LLMs per [6] and references therein.
[4] As we will discuss, DeepSeek has not exactly open sourced everything [20, 21], but now Hugging Face for example, has cloned them in 24 hours, using Llama [33], as they intended, but probably faster than expected. The value of open sourcing GenAI is discussed in [6, 24], and references therein.

One could wonder if the sequence and timing of the DeepSeek releases also served Chinese political agendas, including countering the Stargate Project.

Regardless of the point of view, it is clear that DeepSeek's success has delivered a significant blow to the confidence of many.

1.1.4 Our Thesis

In this paper, we argue that the prevailing "panic" surrounding DeepSeek is largely irrational and unwarranted, reflecting a short-sighted perspective and a lack of understanding of the broader AI history and development landscape.

Indeed, while the work done by DeepSeek is exceptional, as even admitted for example by OpenAI CEO [9], it is a direct applications of many known and proven principle along for example reinforcement learning, model distillation, lower precision digitization, integerization (quantization)/lower precision, and pruning, also now known as sparsity.

These techniques can be used by anybody, especially with an open sourced approach and model [6,20,21,33]. Many have already undertaken initiatives to build open source clones, match the results, reconstruct the history of the data sources, or understand its systems' prompts and other settings [20,21,33].

The LLMs incumbents, especially OpenAI, have started to counter with promises of better models, new models [20,21,33-36,39,40,48], certainly following the approaches of DeepSeek, and adjusting their prices [35-37,48]. DeepSeek is made available for free by Microsoft copilot [35,36]. There is no doubt that soon a (slew of) better model(s) will appear from incumbents or other startup and/or open source projects. In fact, this may have already occurred in China with at least one other challenger to OpenAI—Kimi K15 [38]—even if it does not surpass DeepSeek. This cycle of progress is inherent to the evolution of GenAI and LLMs, a reality that may have gone unnoticed by some.

These initiatives will also benefit NVIDIA and infrastructure cloud provider by creating new sources of demand [42] due to decreasing costs. Indeed, reducing power consumption, computational requirements, and the cost of training and running models will enable all players to develop more complex algorithms. However, this does not necessarily bring them closer to achieving AGI [6]. Additionally, it will drive increased demand for hardware and computing resources. Moreover, reducing the power consumption, even if temporary, is beneficial for the planet [5,6,43-46].

The geopolitical implications are harder to ignore, but again the advances introduced by DeepSeek are no surprises—much like the substantial number of AI students and experts globally with ties to China.. DeepSeek could have happened anywhere but the fact that it occurred in China shouldn't be a surprise. The next innovation could come from anywhere as well, China, USA, or elsewhere. Of course, DeepSeek being Chinese raises many concerns about privacy, user data storage, IP stealing, unethical practices, censure, government agenda and control, etc. [47,49,52-54]. Many have already questioned how their LLM system was build, possibly using OpenAI and others LLMs as basis for distillation, violating the terms of services of these models [50,51]. Others have pointed out quality issues [56-64], including the possibility that performances on some tests do not provide a high-quality experience for end users, or that content can be damaging and usage creating a security risk. As mentioned, censorship is reportedly encountered on many topics sensitive to China [47]. Large amount of data also seems to be questionably sent to Chinese servers, raising security, privacy and regulatory issues [49].

So, the cost of the DeepSeek models may be much higher than claimed, if one compute the cost of the original models that have been distillated [65,66], and that also applies to hardware. The final step alone cost approximately USD 5M. However, at least the incumbents now have access to these models.

Cheating on performance is hardly a new phenomenon in this space. After all, OpenAI was recently caught red-handed for manipulating its way to a high math/reasoning benchmark score—a scandal that, surprisingly, did not receive much attention. [67,68].

Additionally, we argue that LLMs and generative AI have become commodities, in contrast to AI agents and applications. The monetary value will shift toward all sorts of applications and AI agents, where it would truly matter. While investors may need to grasp this distinction, it is a common challenge with many emerging technologies in addition to AI.

1.2 Research Objectives and Scope

This paper explores the evolving landscape of AI development through historical patterns and recent advancements, particularly focusing on the impact of DeepSeek's latest models. The core research objectives are:

- Analyzing AI Development Through Historical Patterns [69,70]: By examining past trends in AI innovation, we contextualize the rise of new competitors and assess whether their advancements mark a fundamental shift or a continuation of established cycles.
- Evaluating DeepSeek's Technological Impact [71,72]: This paper also examines DeepSeek-R1 and R1-Zero within the broader AI ecosystem, assessing their significance in terms of computational efficiency, reasoning capabilities, and potential influence on global AI competition.

1.2.1 The Competitive Landscape and Technological Implications

The recent release of DeepSeek-R1 and R1-Zero by the Chinese startup DeepSeek has sparked considerable debate in the AI community. These models, particularly R1-Zero, have demonstrated impressive reasoning capabilities, approaching the performance of models from leading U.S. companies such as OpenAI—despite DeepSeek's relatively constrained resources [73,74]. This development raises critical questions about the competitive dynamics of AI research and its implications for U.S. technological leadership.

While DeepSeek's advancements are notable, they do not necessarily signify a decline in U.S. dominance in AI. The competitive nature of Large Language Models (LLMs) is shaped by rapid innovation, the widespread dissemination of research breakthroughs, and increasingly efficient techniques such as model distillation and reinforcement learning. These factors enable emerging players to achieve high performance even with fewer resources, contributing to an ongoing cycle of improvement where industry leaders continuously refine and advance their models in response to new challengers.

1.2.2 Limitations of Current LLM Approaches

Beyond the emergence of DeepSeek, this paper also highlights fundamental constraints within current LLM development. Two key concerns are [75,76]:

1. Finite High-Quality Training Data: As AI models grow in scale, the availability of high-quality, diverse training data becomes a critical bottleneck.

Risks of Model Collapse: The increasing reliance on synthetic data poses challenges, including the potential degradation of model performance over time due to compounding errors [6,25].

1.2.3 The Need for a Broader AI Research Agenda

DeepSeek's progress underscores the importance of sustained investment in fundamental AI research, fostering open collaboration, and exploring alternative architectures beyond LLM-centric models [77,78]. Initiatives like the Stargate Project demonstrate the potential of large-scale AI infrastructure investments, but long-term innovation would also require diversification in research directions, including multi-agent systems, cognitive architectures, and adaptive AI frameworks.

Maintaining a balanced perspective is crucial. While new breakthroughs—such as DeepSeek's models—may challenge existing paradigms, they should be viewed within the broader trajectory of AI evolution. The key to continued leadership in AI will not be reacting to individual advancements but fostering an ecosystem that prioritizes innovation, strategic investment, and cross-disciplinary research.

1.2.4 Disclaimer

This paper presents a technical analysis with considerations in business, investment, and geopolitics; however, it does not claim to predict with certainty the actual impact of DeepSeek's advancements, the Stargate joint venture, or the broader AI strategies of the United States and China. AI innovation—both in models and other areas—is evolving at such a rapid pace that it is impossible to keep up with the latest developments, despite our best efforts. The long-term implications remain uncertain—ranging from negligible to transformative—but we argue that such advancements [37] would have occurred regardless and that their significance should be viewed in the context of broader global reflections on AI development. This analysis is based on current trends and publicly available information, representing one perspective in a rapidly evolving field.

2 DeepSeek's Technological Advancements

2.1 DeepSeek V3

The DeepSeek V3 model [22] introduces further enhancements that consolidate its position in the market. Key improvements in DeepSeek V3 include [20] Open source collaboration. By adopting an open source approach,

DeepSeek V3 encourages community contributions and accelerates innovation. The availability of open source code and methodologies fosters a collaborative environment where researchers and developers can build upon existing work.

2.2 Key Innovations in DeepSeek-R1 and R1-Zero, and Future Directions

The work done by DeepSeek, with DeepSeek-R1 and related models [23], is impressive. They have managed to develop a top LLM model and system on limited AI hardware and at a very low cost. Traditionally, developing top LLMs like GPT-4 / Llama has been prohibitively expensive, often exceeding $100 million or more and requiring tens of thousands of high-end GPUs. This cost restricts AI innovation to only the largest corporations, or super funded unicorns. However, see our previous comments that cost of the DeepSeek models are actually way higher, as argued in section 1.1.4.

DeepSeek R1 showcases several technological advancements that position them as a competitive player in the AI landscape. Noteworthy innovations include [23]:

- Numerical Precision Reduction: By reducing numerical precision from 32 to 8 decimal places, DeepSeek achieves a 75% decrease in memory usage with minimal performance impact. This innovation is particularly significant as it allows for efficient utilization of resources, making high-performance models more accessible.
- Memory Optimization: Advanced memory optimization techniques enable DeepSeek to operate on standard/lower grade AI hardware. This breakthrough reduces the dependency on expensive and specialized GPUs, democratizing access to cutting-edge AI technology.
- Multi-token Reading: DeepSeek's ability to process text in chunks, rather than word-by-word, effectively doubles the processing speed while maintaining accuracy. This innovation enhances the efficiency of language model operations, fostering faster and more reliable outputs.
- Selective Parameter Activation: With a parametric framework that activates only 37 billion out of 671 billion total parameters as needed, DeepSeek minimizes computational overhead. This selective activation ensures that the system remains highly efficient and cost-effective.

These advancements collectively reduce the cost of training AI models from approximately $100 million to around $5 million and cut down the required hardware from 100,000 GPUs to a manageable 2,000 [2,37].

DeepSeek also released a set of distilled models, for example from Llama [23]. These models, used by DeepSeek for later projects could then be even more efficient that the original DeepSeek-R1.

DeepSeek R1 is accompanied by an Enhanced Image Generation feature: the new image generator, Janus, integrates sophisticated algorithms to produce high-quality images with remarkable detail and realism [8,9]. This feature extends DeepSeek's capabilities beyond text processing, opening new capabilities for creative applications.

2.3 Open Source

DeepSeek is open source: it means their source code is publicly available for anyone to view, use, modify, and share. By making their code open source and methodologies publicly available, they lower the barriers to entry, allowing a large community to compete without requiring billion-dollar budgets. It should help collaboration and

accelerate innovation across the industry, and allow DeepSeek to benefit from contributions and ideas that that community generates.

However, note that, while DeepSeek is deemed open source, it is not necessarily transparent. There is a lot of information missing about how the models have been built, with what data and with what tools [21]. It's possible that the data they used wasn't entirely legitimate. [50,51]. This is also in part why others are not only trying to reproduce DeepSeek, but also to figure out the public data set that it requires, or the system prompts that then manage the overall offerings [20,21,33]. In that sense rebuilding and building DeepSeek from Llama (distillation) leads to variations of DeepSeek-R1, not the original R1. All the literature and research can freely mix and match the actual DeepSeek-R1 they study.. Detail technical works need to trace back the exact genealogy of each DeepSeek model used. Then yet again, as we already explained, all LLMs behave roughly the same when using the same algorithm to be trained at construction and at execution.

However, DeepSeek embraced the open-source movement, championed by Meta with Llama. It matters as we will further discuss later [24]. Open source can be argued as a key factor in DeepSeek success: they managed to access (legally) models, code and dataset on that basis. The jury is still out on what else may have occurred, as we'll discuss later.

2.4 Learning From the Efficiencies Applied by DeepSeek

The implications of DeepSeek innovations are profound. By reducing costs and lowering barriers to entry, they have paved the way for broader participation in AI research and development. The open source model promotes transparency and inclusivity, driving collective progress in the field.

The efforts and innovations, exemplified by DeepSeek, underscore the importance of continued investment in AI research, fostering a collaborative ecosystem, and exploring diverse technological directions. While DeepSeek's advancements are remarkable, the broader context of AI development remains dynamic, with numerous players contributing to the evolution of the field.

3. LLMs: Easy To Reproduce, Match or Surpass

3.1 Towards Commoditization Among a Generation of LLMs

It turns out that LLMs are easy to reproduce, match or improve and even surpass. DeepSeek showed the world that it can be done as a commercialization exercise and not just a research exercise [37]. But past previous research and papers had already demonstrated this..

Optimization for a certain target, domain, or use cases can be done for example with retraining (a very expensive option), fine tuning, reinforcement learning, RAG, or knowledge distillation. This list is not exhaustive and each approach has many variations.

It is also worth noting that once a certain size (number of parameters) is considered, data scope, and underlying technologies typically published, or open sourced, most top-tier LLMs of that generation perform at a same or similar level, with little change in results or performance when switching one for another[5] [81][6].

Open source and data sets are key to ensure that this continues.

As almost all publicly available data has been used for training differences will be further reduced [6].

In this section, we present a few examples to illustrate this..

3.1.1 Understanding LLMs: Training, Fine-Tuning, and Inference

At the heart of the modern AI systems are models, mathematical structures designed to recognize patterns in vast amounts of data. In the context of LLMs, a model is an artificial neural network trained to process and generate human-like text. These models, such as GPT-4, DeepSeek-R1, or LLaMA, learn by analyzing billions of words, developing an understanding of language structures, meanings, and even reasoning patterns.

Training: Building the Foundation

Training a model is akin to teaching a new employee everything from scratch. This process involves feeding the model massive datasets—ranging from books and articles to code repositories—while it continuously adjusts billions (or even trillions) of parameters to improve its understanding. The training process is computationally expensive, requiring powerful hardware like GPUs (Graphics Processing Units: processors designed for handling graphics and parallel computations, commonly used in gaming and AI) or TPUs (Tensor Processing Units: Google's specialized chip for accelerating machine learning and AI tasks) and significant time investment.

Training is divided into two stages:

1. Pretraining: The model learns general language patterns by predicting missing words in a sentence (self-supervised learning). At this stage, it gains broad knowledge but lacks specific expertise.
2. Supervised Fine-Tuning (Optional): Developers refine the model by training it on specific tasks using curated datasets with labeled answers, enhancing accuracy in targeted domains.

Fine-Tuning and Super Fine-Tuning: Adapting to Specialized Needs

Once a model is pretrained, it can be fine-tuned for specific applications. Fine-tuning is the equivalent of taking a generalist and turning them into an expert. For example, a base LLM may have a general understanding of finance,

[5] For example, the readers can easily check themselves if using perplexity.ai PRO [80] and asking the same search or request switching between the models.

[6] Again, that statement is for LLMs using similar algorithms, and possibly sizes (not always as we see SMLs matching ChatGPT [16]) etc. [81], otherwise performance differ a lot from generation to generation. Also this may not be true for applications requiring a use case or domain specific model. There the data used will matter a lot. It is detailed in the next subsections.

but fine-tuning it with SEC filings, market reports, and investment strategies enhances its proficiency in financial analytics.

There are two primary approaches to fine-tuning:

1. Regular Fine-Tuning: Adjusts a subset of the model's parameters based on new domain-specific data.
2. Super Fine-Tuning: A more advanced process, often requiring reinforcement learning or instruction tuning. This level of tuning is seen in models like GPT-4-turbo, which optimize for efficiency, accuracy, and alignment with human preferences.

Fine-tuning is required for businesses looking to tailor AI solutions to their industry expertise, compliance needs, and proprietary datasets.

Inference: Transforming AI Knowledge into Action

Once trained and fine-tuned, the model enters inference mode—the phase where it generates responses based on user queries. Inference is where AI creates real business value, from answering customer questions and writing reports to coding and making complex predictions. Unlike training, which is a one-time or periodic computationally heavy process, inference happens in real-time and is optimized for efficiency.

Companies deploying AI LLMs must balance accuracy, speed, and cost in inference [113]. Optimized inference reduces latency (response time) and lowers computational expenses, which is why cutting-edge research focuses on more efficient model architectures, quantization techniques, and hardware acceleration (enhancing performance by using specialized hardware to offload and speed up specific tasks).

3.1.2 Why This Matters for Businesses and C-Suite Executives

For CEOs, CTOs, and decision-makers, understanding these AI processes is critical for making informed strategic choices:

- Training determines a model's foundational intelligence but is expensive and resource-intensive.
- Fine-Tuning allows companies to create AI solutions tailored to their needs, increasing accuracy and relevance.
- Inference is where AI interacts with customers, employees, and systems, requiring cost-effective deployment strategies.

The rise of compute-efficient models like DeepSeek-R1-Zero demonstrates that raw computational power is no longer the sole advantage—algorithmic efficiency and strategic data utilization play equally critical roles. Companies must decide whether to train their own models, fine-tune existing ones, or use third-party solutions, carefully balancing innovation, cost, and infrastructure capabilities.

By understanding these AI fundamentals, business leaders can navigate the rapidly evolving AI landscape more effectively, using LLMs to gain a competitive edge and drive industry transformation.

3.2 Example 1: Meta’s Llama vs OpenAI’s ChatGPT

Llama Index and variations are open source LLMs allowing low-cost usage when deploying your own, you only need to pay for the computing resources, making it a preferred option for open-source projects and academic/research activities. Llama, Index and its variations, represent an open source LLM that facilitates low-cost usage by allowing users to deploy their own models, thus only incurring expenses for computing resources. This makes it an ideal choice for open-source projects and academic research, promoting accessibility and innovation within the AI community.

Llama, short for Large Language Model Meta AI, was first introduced in a research paper published in February 2023 [10]. This initial release comprised a collection of foundation language models with varying parameter sizes, ranging from 7 billion to 65 billion. These models were trained on a massive dataset of text and code, drawing from publicly available sources such as CommonCrawl, C4, GitHub, Wikipedia, books, ArXiv, and StackExchange [11].

A key contribution of the Llama research was proving that the state-of-the-art LLMs could be trained effectively using exclusively publicly available datasets, without relying on proprietary or restricted data [10]. This approach fostered open access and facilitated further research and development within the AI community, democratizing access to LLMs for researchers who may not have the resources to train such large models from scratch [12].

Unlike models like GPT-3, Llama utilizes the SwiGLU activation function instead of GeLU, rotary positional embeddings (RoPE) instead of absolute positional embeddings, and RMSNorm instead of layer normalization [13]. These architectural choices contribute to Llama's efficiency and performance.

Training a ChatGPT-like LLM, including Llama, typically involves two stages as discussed earlier: pre-training and fine-tuning,. In the pre-training stage, the model is exposed to a massive amount of text data to learn general language patterns and relationships. This is followed by the fine-tuning stage where the model is trained on a more specific dataset, often with human feedback, to refine its behavior and align it with the desired task, such as conversational response generation.

Llama 3, released in April 2024, expanded the capabilities of LLMs by incorporating multimodal functionalities. Through a compositional approach, Llama 3 integrated image, video, and speech processing capabilities [15]. This advancement allows the model to process and understand information beyond text, expanding its potential applications in areas like image recognition, video analysis, and speech processing.

Llama offers opensource LLMs (and SMLs) with performance and capabilities that rival major proprietary LLM providers—at a significantly lower cost by design.

Llama's open-source structure, optimized computational requirements, and potential for SMLs make it a viable choice for running derived models locally.

3.3 ChatGPT o1 Match by Distillation

In machine learning, knowledge distillation (or model distillation) is the process of transferring knowledge from a large model to a smaller one. While large models—such as deep neural networks or ensembles—have greater capacity, they often do not fully utilize it [17]. Evaluating these models remains computationally expensive, even when much of their capacity goes unused.

Knowledge distillation enables the extraction and transfer of essential knowledge from a large model to a smaller one, preserving performance while significantly reducing computational requirements. Since smaller models are more efficient, they can be deployed on resource-constrained hardware, such as mobile devices. This technique has been in use for over a decade, but recent research has demonstrated its effectiveness in producing small language models (SLMs) with performance comparable to large language models (LLMs), such as ChatGPT o1.

Before DeepSeek, leveraged model distillation—alongside other well-known optimization techniques, such as lower-precision computation, quantization (integerization), and dynamic neural network pruning—to develop smaller, high-performing models at reduced costs, driving innovation in the field, others had already used distillation to reproduce earlier ChatGPT models (e.g., [16] and references therein).

3.4 ChatGPT o1 Reproduction with Reinforcement Learning

In 2024, it was demonstrated that reinforcement learning [18,19] could be used—alongside knowledge distillation—to build an LLM system with performance comparable to OpenAI's ChatGPT o1 [18].

This work was not led by DeepSeek but leveraged a well-established approach, as reinforcement learning has long been recognized as a key technique in training LLMs. In fact, reinforcement learning played a crucial role in the development of o1 and other large language models, particularly through methods such as reinforcement learning from human feedback (RLHF) and related optimization strategies.

3.5 Model Collapse with Synthetic Data

It is well established that training models on synthetic data—especially when generated by LLMs—can lead to model collapse, as discussed in [6,25] and references therein. OpenAI has reportedly used synthetic data for training o3 [27,30], which could have significant consequences, as has been anecdotally reported online.

Meanwhile, it has been widely observed that most LLMs today are trained primarily on publicly available digitized data, such as content collected (scraped) from the internet [6, 31] (and references therein), with some additional proprietary datasets. As almost all the digitized / publicly available content has been processed, this largely explains why modern LLMs exhibit similar capabilities and behaviors across most tasks: they rely on overlapping training data sources.

As a result, meaningful differentiations[7] now primarily arises from domain-specific training, fine-tuning, or customization tailored to specific use cases, industries, or organizations.

3.6 Countering DeepSeek

[7] It is possible that in the future, efforts to digitize books and other material not current digitized or not available due to copyright considerations, may provide an extra boost, or at least significantly improve the quality of the resulting LLMs. We will discuss this in an upcoming paper.

In response to the rapid advancements in AI, particularly the emergence of cost-effective models like DeepSeek, OpenAI has announced plans to accelerate the release of upcoming models. Collaborating with Microsoft, they intend to offer some of their latest top-tier models for free, albeit with certain restrictions. For instance, Microsoft has made OpenAI's o1 reasoning model freely available to all Microsoft Copilot users through the "Think Deeper" feature [35,36,40,48].

Concurrently [20,21,33,39], Hugging Face and others are actively working to replicate DeepSeek's R1 model, aiming to provide open source access to similar capabilities. These initiatives include efforts to reproduce the full DeepSeek R1 data and training pipeline, ensuring that the AI community can benefit from these advancements.

Future releases of LLMs will learn from DeepSeek models and build on it [37].

These developments underscore the dynamic and iterative nature of the generative AI landscape. As organizations build on each other's innovations, the GenAI field continues to evolve, leading to more accessible and advanced AI solutions. This is the circle of life for progress in the GenAI world. We argue that it is quite normal and expected and inherent to the nature of GenAI.

4. No Real Surprise

4.1. The Industry Already Knew Much of What DeepSeek Did

Considering Sections 2 and 3, it should be clear that DeepSeek effectively combined and applied established GenAI industry techniques, rather than pioneering entirely new methodologies. This isn't a case of Chinese AI surpassing U.S. AI; rather, it underscores how assembling a talented team of engineers and scientists, fostering creativity, and working within budget and hardware constraints can drive meaningful innovation. Realistically, DeepSeek's advancements could have emerged from anywhere, however, China was a prime candidate due to its mix of technical expertise and external constraints that sparked the pursuit of alternative solutions. Much of what DeepSeek implemented was already well understood in the field:

- Knowledge distillation and reinforcement learning: These techniques have been widely used for years and were already applied by other AI leaders in their best LLMs [16,17].
- Reducing parameter precision [82,84]: Lowering precision is a common practice in computer engineering, and it has already been applied to LLMs to optimize performance and efficiency. It can be done without or with limited impact on performance of the resulting LLMs.
- Sparsity [89,90] and pruning: The idea of reducing model complexity through pruning dates back to the 1990s, when it was applied to decision trees, maximum likelihood algorithms, and early neural networks [85,88].

DeepSeek's success lies in their ability to integrate existing techniques into a cost-effective LLM framework, enabling high-performance models with efficient, low-cost training.

4.2 Credible?

However, some have raised concerns that DeepSeek may have leveraged unauthorized access or distillation of models from OpenAI and other LLM providers as a starting point [50,51]. If this were the case, the reported cost savings would not account for the original cost of developing these models in the first place [65,66].

So, in reality, the overall cost of training DeepSeek's models is not drastically lower than that of other LLMs. Instead, the key cost reduction appears to come from efficiency optimizations and refinements applied on top of existing techniques.

4.3. Geopolitical Concerns About DeepSeek Models

DeepSeek methods and success are raising many geopolitical questions, and concerns (as referenced earlier in this paper):

1. Potential Unauthorized Model Distillation
    - There are allegations that DeepSeek may have trained its models using unauthorized access or distilled outputs from existing OpenAI and other Western LLMs. If true, this raises concerns about intellectual property theft and fair competition in AI development.
2. China's AI Advancements & Strategic Implications
    - DeepSeek's rapid progress demonstrates China's growing AI capabilities, which could challenge U.S. dominance in AI research and innovation. This could accelerate AI arms races and national security concerns in both civilian and military applications.
3. Government Control & Censorship
    - As a China-based AI model, DeepSeek is subject to Chinese government regulations that require alignment with state policies. This raises concerns about censorship, bias, and lack of transparency, particularly when the model is deployed in global applications.
4. Data Security & Privacy Risks
    - If DeepSeek models are deployed internationally, particularly in Western businesses, academia, or government sectors, there may be risks of data leakage, espionage, or surveillance due to China's strict cybersecurity laws, which require state access to data.
5. Restricted Export & U.S. Sanctions
    - The U.S. has already restricted AI chip exports to China, limiting access to advanced NVIDIA GPUs. If DeepSeek models prove competitive, policymakers may consider further trade restrictions to curb China's AI advancements.
6. Disinformation & Influence Operations
    - AI-powered chatbots, content generation, and social media automation could be exploited for disinformation campaigns or influence operations aligned with China's geopolitical interests, similar to past concerns about Russian and Chinese online influence.

7. Global AI Regulations & Ethical Standards
    - DeepSeek operates under a different regulatory framework than US and European AI companies, meaning its safety measures, bias mitigation, and ethical AI standards may not align with those of OpenAI, Google DeepMind, Anthropic, IBM, and others. This complicates efforts for international AI governance.
8. Competitive Pressure on Western AI Firms
    - DeepSeek's emergence as a high-quality, low-cost AI provider could push OpenAI, Google, and Meta to accelerate commercial AI deployments, potentially leading to hasty releases without adequate safety guardrails.

However, it must be noted that competition in the AI landscape, particularly in the realm of large language models (LLMs) and generative AI, is a vital catalyst for innovation and progress. When multiple entities—ranging from established tech giants to emerging startups—compete in this space, it creates an environment where continuous improvement is both necessary and beneficial. Here are several reasons why competition is advantageous in this context:

1. Accelerated Innovation:
    - In a competitive market, companies are constantly driven to push the boundaries of technology. This race to outperform rivals results in rapid advancements, as each participant seeks to offer more efficient, accurate, and versatile AI solutions. The iterative process of research and development is expedited, leading to breakthroughs that might not occur in a monopolistic or complacent environment.
2. Improved Quality and Performance:
    - Competition encourages organizations to enhance the quality of their products and services. For AI models, this means not only better performance metrics—such as reduced error rates and faster processing times—but also more robust safety features and ethical safeguards. As companies vie for market share, they are incentivized to address shortcomings and innovate in areas like model transparency, bias mitigation, and data security.
3. Cost Efficiency and Accessibility:
    - Competitive pressures often lead to cost reductions, making advanced AI technologies more accessible to a broader range of users. This democratization of AI can spur further innovation by allowing academic institutions, startups, and smaller enterprises to experiment and build upon state-of-the-art models without the prohibitive costs typically associated with proprietary systems.
4. Diverse Perspectives and Collaborative Synergies:
    - A competitive ecosystem brings together diverse perspectives and expertise. The interplay between various players—each with their unique approaches and strengths—can foster collaborative synergies, even among competitors. Shared challenges often lead to cooperative efforts in establishing industry standards and best practices, which in turn benefit the broader AI community and society at large.
5. Resilience and Adaptability:
    - Competition forces companies to remain agile and responsive to market needs. In the rapidly evolving field of AI, this adaptability is crucial for developing systems that can cope with emerging challenges,

such as ethical dilemmas, data privacy concerns, and cybersecurity threats. A competitive market helps ensure that AI solutions remain resilient, continuously evolving to meet both technological and societal demands.

6. Regulatory and Ethical Advancement:
    - The presence of multiple players in the market also encourages more robust regulatory and ethical frameworks. When companies compete not only on technological prowess but also on ethical and responsible AI practices, it creates an incentive to develop and adhere to higher standards. This can lead to more transparent, accountable, and ethically aligned AI systems that better serve public interests.

The two sides contrasted above could balance each other.

5. LMMs Cycle of Life: Only Temporary Dominance

5.1. The GenAI Circle of Progress (or Life)

As we have observed and predicted, it is entirely natural for LLMs developed by one vendor or research team—especially when their approaches, code, or data sources are published or hinted at—to be approximated, matched, or even surpassed by newer models, sometimes from the same team. What happened with DeepSeek is business as usual, aside from the potential for cheating and other geopolitical considerations.

New techniques emerge incrementally, whether in model architecture, fine-tuning strategies, reinforcement learning, knowledge distillation, architecture/composition/orchestration of systems [91-94], or the cost-efficient methods pioneered by DeepSeek. These innovations are quickly understood, replicated, and improved upon by others in both open source and proprietary AI projects [33].

As a result, just as DeepSeek's models and approaches, just as the ones from OpenAI and others, are now being analyzed and integrated elsewhere [33,37], future models will follow the same trajectory. New models will emerge—leveraging different methodologies, larger datasets, or building upon existing models—and another leader[8] will take the lead.

This is the cycle of innovation and dominance in generative AI.

Recent developments underscore this reality, such as:

- OpenAI's[9] commitment to releasing more powerful language models [34,35,39,40].
- Microsoft's[10] decision to make o1 freely available to Windows users [35,36,40,48].

The generative AI landscape is continually evolving, with dominance being temporary and innovation relentless.

[8] Or for a while the same leader with new release and generations of LLMs.
[9] And others [39].
[10] And others.

5.2 LLMs Have Plateaued, But They Probably Haven't Peaked

While LLMs may have plateaued in terms of consuming most of the publicly available digitized data [6], this does not mean that their capabilities, or the capabilities of systems/applications/agents using them have reached their peak.

Recent advancements—such as DeepSeek's cost-efficient models and the emergence of reasoning models incorporating techniques like chain-of-thought prompting and its variations—demonstrate that there are still significant opportunities to enhance and extend LLM capabilities.

Lower-cost and faster training methods are opening the door for:

- More complex architectures and algorithms to improve reasoning and adaptability.
- Revisiting and repurposing training data to extract new insights.
- Specialized AI hardware[11], and cloud computing innovations to support larger-scale training and inference.

As these trends continue, the demand for high-performance computing resources, specialized AI chips, and large-scale cloud infrastructure is expected to increase rather than decline. The era of AI expansion is far from over.

6. Open source AI: Potential and Limitations

We already discussed the value of open sourced LLMs (and data set/training tools) and arguments as in [24,26]. Open source AI supporter see it as the only true way forward.

So, in addition to section 2.5, we expand the open source analysis in the following sub sections.

6.1 Advantages of Open Source AI

Open source AI offers numerous advantages, particularly in the realms of technical transparency, collaborative development, and research accessibility. One of the primary benefits is the ability to conduct detailed security audits, assess potential biases, and verify ethical considerations. The open availability of code and model architectures allows researchers and developers to scrutinize and refine AI systems, ensuring that they adhere to ethical standards and security best practices [95,96].

Collaboration is another cornerstone of open source AI. A community-driven approach fosters innovation by leveraging distributed expertise, enabling rapid iterative improvements, and incorporating diverse perspectives

[11] Although, to be honest, if we observe the evolution of computing so far, specialized CPU have often be replaced by more generic platforms with generic CPU. That is still an evolution we expect to see happen with just CPU at some point in the future including GPU capabilities or modules. But it may end up being just an academic argument, as we admit that CPU will then have also evolved to look in part like GPUs.

from global contributors. This decentralized model accelerates the development cycle and enhances the robustness of AI solutions [98,100].

From a research and innovation standpoint, open source AI lowers barriers to entry, reducing both computational and financial constraints. This democratization of AI technology empowers academic institutions, startups, and independent researchers to experiment with and build upon existing models without the prohibitive costs associated with proprietary solutions. Additionally, the flexibility of open source frameworks allows developers to fine-tune models for specific applications, customize architectures to suit unique needs, and implement targeted modifications that drive domain-specific advancements [96,99].

6.2 Limitations and Challenges

Despite its advantages, open source AI faces several challenges. Security risks can arise from the code's public accessibility, potentially exposing vulnerabilities that could be exploited [95,97]. But as is well known and proven with traditional software, exposing the code, models and training to public scrutiny and many pairs of eyes typically ensure quick discovery of problems and risks, and faster remediations[12].

The lack of official support can pose challenges in critical situations, making it advisable to have a troubleshooting plan or work with an AI partner [95].

While access to open source models is often free, there are additional costs associated with deploying and maintaining them [95,96].

Open source AI offers numerous advantages, particularly in the realms of technical transparency, collaborative development, and research accessibility. One of the primary benefits is the ability to conduct detailed security audits, assess potential biases, and verify ethical considerations. The open availability of code and model architectures allows researchers and developers to scrutinize and refine AI systems, ensuring that they adhere to ethical standards and security best practices [95,96].

Collaboration is another cornerstone of open source AI. A community-driven approach fosters innovation by leveraging distributed expertise, enabling rapid iterative improvements, and incorporating diverse perspectives from global contributors. This decentralized model accelerates the development cycle and enhances the robustness of AI solutions [97,98].

From a research and innovation standpoint, open source AI lowers barriers to entry, reducing both computational and financial constraints. This democratization of AI technology empowers academic institutions, startups, and independent researchers to experiment with and build upon existing models without the prohibitive costs associated with proprietary solutions. Additionally, the flexibility of open source frameworks allows developers to fine-tune models for specific applications, customize architectures to suit unique needs, and implement targeted modifications that drive domain-specific advancements [99,100].

[12] That is true as long that the community of support / development has active members, and that means typically a large enough community of developers and users. Otherwise, nobody may end up being there to fix the protocols. As an example, we have seen this happen with OpenSSL libraries in the past [114].

6.3 Data Source Considerations

The integrity of an AI model is deeply tied to the quality and reliability of its training data. Platforms such as the Hugging Face ecosystem have become instrumental in providing extensive repositories of open source AI models and datasets. Hugging Face's Model Hub, dataset repository, and collaborative research spaces serve as essential resources for researchers and developers seeking high-quality AI tools [98,100].

Verification mechanisms within these platforms play a crucial role in mitigating risks associated with unreliable models. Community ratings, usage statistics, and detailed model card documentation help users assess the credibility and applicability of various models before deploying them in real-world applications. These verification methods contribute to transparency and promote responsible AI development within the open source ecosystem [99,100].

6.4 Mitigation of Potential Risks

To ensure the responsible development and deployment of open source AI, comprehensive vetting procedures are essential. Rigorous model evaluations, ethical use guidelines, and transparent documentation can help establish trust and reliability in AI systems. Community governance further strengthens oversight by enabling collective responsibility, facilitating rapid vulnerability identification, and fostering a culture of accountability [98,101].

Preventing the reprehensible use of AI requires proactive security measures. Implementing access controls, such as responsible use policies and technical restrictions on high-risk applications, helps mitigate potential misuse. Ethical development frameworks should include clear usage guidelines, proactive strategies to prevent harmful applications, and continuous monitoring mechanisms to identify emerging risks [97,101].

While open source AI offers transformative potential, its long-term success depends on striking a balance between accessibility and responsible governance. By fostering collaboration, ensuring transparency, and addressing ethical concerns, the open source AI community can continue to drive innovation while mitigating risks associated with unrestricted AI development [98,101].

7. Large LLM Providers: Premature Obituary

7.1 Current Market Dominance

The landscape of large language models (LLMs) is currently dominated by large proprietary AI providers, largely due to their significant infrastructure investments and vertically integrated ecosystems. These organizations allocate billions of dollars to research and development, leveraging custom-built AI training infrastructure to refine and scale their models. The massive computational resources available to these providers allow for continual model refinement, ensuring that their offerings remain at the forefront of AI capabilities [102].

A key component of their dominance is vertical integration, which enables an end-to-end control of AI ecosystems. From dataset curation to model deployment, these companies manage tightly controlled development pipelines, optimizing performance while maintaining proprietary advantages. Additionally, they implement comprehensive monetization strategies, ensuring sustained revenue generation and continued investment in AI advancements [103].

The competitive strengths of proprietary AI providers stem from their advanced training capabilities and enterprise-grade solutions. Access to vast, curated datasets allows for sophisticated model refinement, while large-scale iterative improvements enhance performance. Furthermore, enterprise clients benefit from robust compliance and security frameworks, predictable performance guarantees, and extensive support infrastructure — elements that reinforce the dominance of these providers in commercial applications [78,104].

7.2 Barriers to Disruption

Challenging the entrenched position of proprietary AI providers presents significant hurdles, particularly in terms of technical and economic barriers. Computational costs for training large-scale models are exponentially high, necessitating specialized hardware and complex infrastructure management. Smaller players often struggle to match the sophisticated training methodologies, advanced architectural innovations, and nuanced performance optimization that industry leaders employ [105].

Beyond technical challenges, economic constraints further deter market disruption. Developing competitive AI models requires billions of dollars in research funding, continuous hardware upgrades, and extensive talent acquisition. The AI landscape is also influenced by strong network effects, as established providers benefit from brand credibility, existing enterprise relationships, and a proven track record of innovation. These factors collectively create formidable barriers that limit the ability of new entrants to compete at scale [106].

7.3 Open Source and Emerging Challenges

Despite the current dominance of proprietary AI, open source and decentralized AI approaches present potential disruptive factors. Community-driven innovation has led to rapid advancements in AI model development, with distributed research efforts producing increasingly sophisticated open source alternatives. Llama was so far the best example [10-15], as discussed in section 3.2. Emerging decentralized AI methodologies, such as blockchain-based AI and federated learning, aim to distribute computational resources, reducing dependency on centralized infrastructures [107].

While these developments introduce competitive pressure, their impact remains incremental rather than transformative. Open source AI models are improving, but they still lack the extensive computational backing and enterprise-grade reliability that proprietary providers offer. However, the continued evolution of decentralized approaches suggests that the competitive landscape may shift in the long term [108].

7.4 Realistic Projection

In the short to medium term, proprietary AI providers may be able to maintain their dominance due to their resource advantages, established infrastructure, and sustained innovation cycles. While open source models will continue to improve, they are unlikely to displace proprietary solutions outright. Instead, hybrid development models may emerge, blending proprietary advancements with open source contributions to create more dynamic AI ecosystems [109].

The competitive landscape is also evolving toward increased specialization, with niche-specific model development and collaborative innovation frameworks gaining traction. However, the probability of immediate disruption remains low. Large LLM providers are far from obsolete, as substantial barriers continue to protect their market position. Nevertheless, the pressure for continuous innovation remains high, necessitating ongoing advancements to maintain a competitive edge [110].

In any case, what happens next will depend also on what will come next after DeepSeek: will the cost remain high because GenAI, now adds new complex algorithms, using the money possibly made available by learning the lessons form DeepSeek, or will we plateau at what we are at lower cost. In the latter case, many new entrants will appear.

7.5 Strategic Implications: The Stargate Joint Venture

The analysis of current market conditions strongly supports the case for collaborative ventures such as the recently proposed Stargate joint venture. Given the significant barriers to entry in AI development, pooling resources among multiple organizations presents a viable strategy for overcoming computational and economic challenges [111].

Strategically, Stargate aligns well with the realities of AI competition. By aggregating computational resources, the venture can mitigate the prohibitive training costs associated with state-of-the-art AI development. Shared infrastructure reduces individual financial risks, while distributed expertise fosters innovation more efficiently than a single entity operating in isolation [112].

The potential advantages of Stargate include the ability to combine the strengths of multiple organizations, leveraging their respective capabilities to create a robust alternative to existing large LLM providers. This collaborative approach directly addresses key challenges identified in this paper:

- **Computational Barriers:** Stargate pools computational resources, overcoming limitations faced by individual entities.
- **Economic Constraints:** A shared investment model reduces financial risk while maintaining competitive AI development efforts.
- **Decentralized AI Approaches:** The venture aligns with emerging trends in collaborative model development, fostering innovation across diverse research teams.

Given these considerations, Stargate Joint Venture represents an innovative and strategic response to the prevailing AI market dynamics. A joint venture approach offers a promising pathway to mitigating the entrenched dominance of proprietary LLM providers, fostering a more diverse and competitive AI ecosystem.

7.6 Our Thesis: The Future of LLM Providers—A Decision for Them to Make

As we have discussed, LLM providers, proprietary providers and open source once sike Meta, have:

- The computing environment/infrastructure
- Data sets
- Training tools
- Skills and resources
- Funding
- Partnerships

With this they have dominated so far.

Our thesis is that DeepSeek has introduced new ways to optimize aspects of LLMs, just as many others. These will render training and execution of LLMs more efficient and cost effective. But LLMs are still far behind the goal of AGI and they still have much to improve, beyond throwing at them more data[13].

Without knowing the future, and considering the real cost behind DeepSeek [50,51,65,66], it is safe to bet that improvement will again be costly in terms of computing resources, data, skills etc. Clearly large LLM providers will still have a leg up, until, or unless, they are disrupted with a dramatically new approach[14], instead of evolutionary progresses, as we have shown that DeepSeek is just evolutionary. DeepSeek and other newcomers e.g., [38], may or may not be able to catch up on enough resources for the next steps. Other may and will and they will disrupt till the dramatic disruption that we hope for.

AI applications and agents may also disrupt a we discuss in the next section. But again, incumbents with their network of partners and ecosystems, are probably in good position, for a while. What may trip them however is the commoditization of LLMs and GenAI that comes along. Some will thrive, and may believe that opensource LLMs will be one of them. The Google and Microsoft of this world have other business, already using GenAI. They will also be able to continue dominating the LLM space, even as LLM are commoditized: they will be at the core of the cognitive services and framework that they provide for developers to build AI apps and agents, and may also be provider of those.

8. Beyond LLMs: The Emerging Landscape of AI Evolution

We argue that LLMs are being commoditized, because of similar performances within a same generation[15] and the evolution to AI applications and AI agents, and not the path to AGI [6] and references therein.

---

[13] With the caveat of our earlier comment about digitizing content today locked in paper books, or protected by stringent copyrights, but that may or may not lead to just marginal progress.
[14] E.g., as needed to reach AGI [6] and references therein.
[15] This is without training, find-tuning / customization for a specific domain, enterprise or set of use cases. There they may matter but would again be equivalent if same training or fine tuning with the same data was applied, and now open source considerations and tooling matter.

8.1 The Commoditization of LLMs and GenAI

The evolution of Large Language Models (LLMs) and Generative AI (GenAI) is following a clear trajectory toward commoditization. This shift is driven by factors such as architectural standardization, API-driven abstraction, and increased market competition. LLMs are increasingly being integrated as interchangeable infrastructure components, where seamless API interfaces allow businesses to switch providers easily, optimizing their AI strategies without vendor lock-in.

A key development in this transition is multi-model orchestration, where AI applications dynamically select LLMs based on cost, performance, and domain specificity. Businesses can integrate multiple models, switching in real-time to balance efficiency and accuracy, which is especially crucial in high-stakes industries like healthcare, finance, and law. This model-agnostic approach is exemplified by platforms like Perplexity.ai Pro, which enables users to select from various LLM providers such as ChatGPT, Claude, Gemini, and Mistral LLMs. Even statically, developers and enterprise can decide at some point to switch from one LLM provider to another. All this is inherent to the circle of life of LLMs. Indeed they evolve so fast, and so far were so costly, that any developer, third party provider, application provider or user will want to be able to pick up the next greatest new one, which may have new APIs for new capabilities and may not be provided by the same vendor, or switch provider considering hard to predict and understand high costs, need to build they applications, agents or workflow so that switching LLM amounts to changing the URL they point to (and may be do some API transformation).

As standardization of APIs or at least capabilities, increases, competition among LLM providers is shifting from model development to strategic deployment and fine-tuning. The rise of model aggregation platforms further erodes individual model differentiation, accelerating commoditization. Ultimately, the long-term value in AI is transitioning from standalone LLMs to their orchestration within broader AI ecosystems.

As, or when, fine tuning and other custom training become necessary, keeping in mind it is not always the case that it is useful, developers and customers will also be naturally driven towards the ones that have the best tools, cheapest cost of training/fine-tuning, data sets/models and community to assist. Open source will probably benefit significantly.

The bottom line is that LLMs are increasingly becoming commoditized in terms of their providers. However, this does not apply across generations. For instance, an LLM from the current generation, such as R1 or o3, will become commoditized within that generation but will still outperform LLMs from previous generations, like V2 or o.

8.2 The Rise of AI Agents and Applications

While LLMs become increasingly commoditized, AI agents and applications are emerging as the primary interface between users and AI systems. These agents act as intelligent orchestrators, determining in real time which LLM to query and how based on task complexity, cost efficiency, and required expertise, and how to recombine.

Rather than relying on a single LLM engine, AI agents employ adaptive routing techniques to optimize responses across multiple models. This architecture enhances workflow automation, decision-making, and personalized interactions. The growing adoption of AI agents signals a shift away from direct LLM usage toward integrated, application-driven solutions. It is a variation on the notions of MultiAI that we have introduced [6,91-94,109]

Note that while AI agent are nothing new. Applications orchestrate AI engines. AI agent do the same, asynchronously on their own, and they may be triggered all at once and with they result aggregated at the end. Ai workflows can do the same sequentially[16], the recombine the results when all have completed [110].

8.3 LLMs Are Not the Path to AGI

Despite their advancements, LLMs are fundamentally inadequate for achieving Artificial General Intelligence (AGI). Their architecture is based on probabilistic token prediction rather than true comprehension, reasoning, or adaptability. Several key limitations prevent them from evolving into AGI systems [6]:

- Lack of causal reasoning: LLMs cannot infer cause-and-effect relationships, making them unreliable for logical deduction.
- Limited world model: Their understanding is constrained by training data, preventing the formation of an evolving, structured knowledge base.
- Hallucination tendencies: LLMs generate misleading information due to overreliance on statistical associations rather than factual accuracy. It is discussed in [94] and references therein.
- No intrinsic intentionality: Unlike human cognition, LLMs lack self-directed goals or independent thought processes.
- Absence of motivation: AGI requires self-driven learning, curiosity, and goal-setting—qualities entirely absent in current LLMs.

For AGI to emerge, AI systems must integrate more advanced reasoning, memory, and adaptive learning capabilities. Future breakthroughs will likely involve entirely new architectures that go beyond statistical pattern recognition and token prediction. See [6] for a more in-depth discussion of our view.

8.4 The Future: AI Agents, AI Workflows and AI Applications Over LLMs

As AI continues to evolve, the real innovation will lie in the orchestration and deployment of AI agents rather than in the models themselves. The ability to seamlessly switch between LLMs and leverage their strengths dynamically will define the next generation of AI-driven applications [102]. With platforms already enabling model-agnostic AI experiences, the dominance of any single LLM is temporary [103]. The real competitive advantage will belong to those who build the most effective AI-driven systems rather than those who develop individual models [104].

And then there is AGI and so many other conventional AI and ML like predictive AI, explainable AI, ethical AI. LLM and GenAI is such a small part of the field, use cases and revolutions that will eventually take place. It is to be seen if there will be another AI winter in between after all the hype ill placed on just LLMs /GenAI.

[16] [110] discusses when to best use on pattern or another.

9. Conclusions

In conclusion, while LLMs have transformed AI applications, their role should be viewed as a steppingstone rather than the final stage of AI evolution. The future of AI development will be driven by innovative integration strategies, multi-agent/multiAI systems, and research into cognitive architectures that push beyond the constraints of current models.

New use cases, new techniques and algorithm will be introduced for broad AI, and for LLMs. Some will overlap between GenAI and AGI or predictive AI, as LLM does today already on stock trading while not being good at predictive AI.

The DeepSeek contributions follow this trend, and the circle of life for LLMs. Putting different technical idea together showed skills, and efficiencies, and shook up the industry. Now everybody else will do the same. But there is really nothing special beyond this. Everybody else will create a next generation LLM build on These contributions, as well as o3 and others. Then others will add new capabilities. It is really to be determined if DeepSeek will be able to continue to compete and contribute new ideas.

Regardless, the demand for next-generation LLMs, applications, and agents will inevitably drive increased need for AI hardware and computing resources. This is good news for cloud providers and AI hardware vendors. Even geopolitical considerations are limited: sanction AI hardware was not used, or not used much and the sanction worked. China has invested heavily in AI research, and it's expected they can innovate at a comparable level to others. However, it appears they have built on existing Western LLM models and may have taken some legal and ethical shortcuts, much like OpenAI seems to have done.

Yet, open source has emerged as a clear winner from these developments, strengthening the case for open data sets and model-building/training tools. Investors, researchers, and technology leaders who recognize this shift will be well-positioned to lead the next wave of AI breakthroughs.

It’s business as usual.

10. Acknowledgements

The Author wants to thank A.J. Cave for the collaboration leading to this paper.

———

References

[1]: John Power, (2025), “What's DeepSeek, China's AI startup sending shockwaves through global tech?”, Al Jazeera, https://www.aljazeera.com/economy/2025/1/28/why-chinas-ai-startup-deepseek-is-sending-shockwaves-through-global-tech, January 28, 2025.

[2]: Wikipedia, “DeepSeek”, https://simple.wikipedia.org/wiki/DeepSeek. Retrieved on February 1, 2025.

[3]: Graham Barlow, (2025), “What is DeepSeek? Everything you need to know about the new ChatGPT rival that's taken the App Store by storm”, TechRadar, https://www.techradar.com/computing/social-media/what-is-deepseek-everything-you-need-to-know-about-the-new-chatgpt-rival-thats-taken-the-app-store-by-storm, January 27, 2025.

[4]: OpenAI, (2025), “Announcing The Stargate Project”, https://openai.com/index/announcing-the-stargate-project/, January 21, 2025.

[5]: Allison Morrow, (2025), “DeepSeek just blew up the AI industry’s narrative that it needs more money and power”, https://www.cnn.com/2025/01/28/business/deepseek-ai-nvidia-nightcap, January 28, 2025.

[6]: Stephane H. Maes, (2024), “The Trouble with GenAI: LLMs are still not any close to AGI. They will never be”, https://zenodo.org/doi/10.5281/zenodo.14567206, https://shmaes.wordpress.com/2024/12/26/the-trouble-with-genai-llms-are-still-not-any-close-to-agi-they-will-never-be/, December 25, 2024. (osf.io/qdaxm/, viXra:2501.0015v1).

[7]: Zeyi Yang, (2025), “How Chinese AI Startup DeepSeek Made a Model that Rivals OpenAI. When Chinese quant hedge fund founder Liang Wenfeng went into AI research, he took 10,000 Nvidia chips and assembled a team of young, ambitious talent. Two years later, DeepSeek exploded on the scene”, Wired, https://www.wired.com/story/deepseek-china-model-ai/, January 25, 2025.

[8]: The Register, “DeepSeek isn't done yet with OpenAI – image-maker Janus Pro is gunning for DALL-E 3”, https://www.theregister.com/2025/01/27/deepseek_image_openai/, January 27, 2025.

[9]: Samantha Kelly, (2025), “DeepSeek Strikes Again With AI Image Generator Janus-Pro That was fast. The Chinese startup's latest AI model takes on image generation”. CNET, https://www.cnet.com/tech/services-and-software/deepseek-strikes-again-with-ai-image-generator-janus-pro. Retrieved on January 27, 2025.

[10]: Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, Guillaume Lample, (2023), “Open and Efficient Foundation Language Models”, arXiv:2302.13971v1.

[11]: Abid Ali Awan, (2024), “Introduction to Meta AI's LLaMA: Empowering AI Innovation”, DataCamp, https://www.datacamp.com/blog/introduction-to-meta-ai-llama, June 6, 2024.

[12]: Meta, (2023), “Introducing LLaMA: A foundational, 65-billion-parameter large language model”, AI at Meta, https://ai.meta.com/blog/large-language-model-llama-meta-ai/, February 24, 2023.

[13]: Wikipedia, “Llama (language model)”, https://en.wikipedia.org/wiki/Llama_(language_model). Retrieved for this paper on February 1, 2025.

[14]: Rachit Narang , (2024),”Understanding Large Language Models (LLMs) like LLaMa by Meta (Part 1)”, & “Understanding Large Language Models (LLMs): Scaling, Applications & The Future (Part 2)”, https://datahacker.rs/understanding-large-language-models-llms-like-llama-by-meta-part-1/, and https://datahacker.rs/understanding-large-language-models-llms-scaling-applications-the-future-part-2/, November 26, 2024,

[15]: Vijay Maurya, (), “Introducing Llama 3.1 : Key points of paper”, Medium, https://medium.com/@vkmauryavk/introducing-llama-3-1-key-points-of-paper-165c29d9c7fd, July 23, 2024.

[16]: Zhen Huang, Haoyang Zou, Xuefeng Li, Yixiu Liu, Yuxiang Zheng, Ethan Chern, Shijie Xia, Yiwei Qin, Weizhe Yuan, Pengfei Liu, (2024), “O1 Replication Journey — Part 2: Surpassing O1-preview through Simple Distillation, Big Progress or Bitter Lesson?”, arXiv:2411.16489v1.

[17]: Wikipedia, “Knowledge distillation”, https://en.wikipedia.org/wiki/Knowledge_distillation. Retrieved on November 28, 2024.

[18]: Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Bo Wang, Shimin Li, Yunhua Zhou, Qipeng Guo, Xuanjing Huang,Xipeng Qiu, (2024), "Scaling of Srearch and Learning: A Roadmap to Reproduce o1 from Reinforcement Learning Perspective", arXiv:2412.14135v1.

[19]: Wikipedia, "Reinforcement learning", https://en.wikipedia.org/wiki/Reinforcement_learning. Retrieved on February 1, 2025.

[20] Kyle Wiggers, (2025), "Hugging Face researchers are trying to build a more open version of DeepSeek's AI 'reasoning' model", TechCrunch, https://techcrunch.com/2025/01/28/hugging-face-researchers-are-trying-to-build-a-more-open-version-of-deepseeks-ai-reasoning-model/, January 28, 2025.

[21]: Paul Hill, (2025), "Hugging Face wants to make DeepSeek R1 fully open by filling closed source gaps", Newwin, https://www.neowin.net/news/hugging-face-wants-to-make-deepseek-r1-fully-open-by-filling-closed-source-gaps/, January 28, 2025.

[22]: DeepSeek-AI, (2024), "DeepSeek-V2 Technical Report", https://github.com/deepseek-ai/DeepSeek-V3/blob/main/DeepSeek_V3.pdf. December 2024.

[23]: DeepSeek-AI, et al. (2025), "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning", arXiv:2501.12948v1.

[24]: Luis E. Romero, (2025), "ChatGPT, DeepSeek, Or Llama? Meta's LeCun Says Open-Source Is The Key", Forbes, https://www.forbes.com/sites/luisromero/2025/01/27/chatgpt-deepseek-or-llama-metas-lecun-says-open-source-is-the-key/, January 27, 2025.

[25]: Ali Borji, (2014), "A Note on Shumailov et al. (2024): `AI Models Collapse When Trained on Recursively Generated Data'", arXiv:2410.12954v2.

[26]: Ben Wodecki, (2025), "Meta's AI chief: DeepSeek proves AI progress isn't about chips", https://www.capacitymedia.com/article/metas-ai-chief-deepseek-proves-ai-progress-isnt-about-chips, January 30, 2025.

[27]: Cogni Down Under, (2024), "The Synthetic Data Revolution: How OpenAI is Reshaping AI Training", Medium, https://medium.com/@cognidownunder/the-synthetic-data-revolution-how-openai-is-reshaping-ai-training-fd47a6f32de4, October 9, 2024.

[28]: Sean Michael Kerner, (2025), "OpenAI o3 explained: Everything you need to know. OpenAI o3 is the successor to the o1 reasoning model. It is the second release from the OpenAI reasoning model branch. The technology was first announced on Dec. 20, 2024.", TechTarget, https://www.techtarget.com/whatis/feature/OpenAI-o3-explained-Everything-you-need-to-know, January 31, 2025.

[29]: Maria Deutscher, (2024), "OpenAI details o3 reasoning model with record-breaking benchmark scores - SiliconANGLE", https://siliconangle.com/2024/12/20/openai-details-o3-reasoning-model-record-breaking-benchmark-scores/, December 20, 2024.

[30]: Dylan Royan Almeida, (2024), "Synthetic data generation (Part 1)", OpenAI Cookbook, https://cookbook.openai.com/examples/sdg1, April 10, 2024.

[31]: Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, Marius Hobbhahn, (2022 & 2024), "Will we run out of data? Limits of LLM scaling based on human-generated data", arXiv:2211.04325v2.

[32]: Latestly, (2025), "ChatGPT Developer OpenAI Likely To Announce AI Super-Agents Capable of Performing Tasks at PhD Level. Sam Altman-run OpenAI is expected to announce advanced AI super-agents with PhD-level

capabilities.", https://www.latestly.com/socially/technology/chatgpt-developer-openai-likely-to-announce-ai-super-agents-capable-of-performing-tasks-at-phd-level-6578606.html. Retrieved on January 20, 2025.

[33]: Benj Edwards, (2025), "Hugging Face clones OpenAI's Deep Research in 24 hours. Open source "Deep Research" project proves that agent frameworks boost AI model capability.", https://arstechnica.com/ai/2025/02/after-24-hour-hackathon-hugging-faces-ai-research-agent-nearly-matches-openais-solution/, February 5, 2025.

[34]: Carl Franzen, (2025), "It's here: OpenAI's o3-mini advanced reasoning model arrives to counter DeepSeek's rise", VentureBeat, https://venturebeat.com/ai/its-here-openais-o3-mini-advanced-reasoning-model-arrives-to-counter-deepseeks-rise/, January 31, 2025.

[35]: Andrew Tarantola, (2025), "ChatGPT's latest model is finally here — and it's free for everyone", DogitalTrends, https://www.digitaltrends.com/computing/openai-officially-releases-o3-mini-reasoning-model/, January 31, 2025.

[36]: Rael Hornby, (2025), "Truly magical" ChatGPT feature comes to Microsoft Copilot — and it's completely free New. By Rael Hornby published 12 hours ago. Forget DeepSeek, Microsoft is bringing the power of OpenAI's o1 reasoning model to Copilot for free", Laptop, https://www.laptopmag.com/ai/microsoft-copilot-think-deeper-features-chatgpt-o1-model-for-free, January 31, 2025.

[37]: Matt Marshall, (2025), "DeepSeek's R1 and OpenAI's Deep Research just redefined AI — RAG, distillation, and custom models will never be the same", VentureBeat, https://venturebeat.com/ai/deepseeks-r1-and-openais-deep-research-just-redefined-ai-rag-distillation-and-custom-models-will-never-be-the-same/, February 6, 2025.

[38]: Mehul Reuben Das, (2025), "Just days after DeepSeek, another Chinese AI company Moonshot launches model Kimi k1.5 that outshines OpenAI", FirstPost, https://www.firstpost.com/tech/just-days-after-deepseek-another-chinese-ai-company-moonshot-launches-model-kimi-k1-5-that-outshines-openai-13857457.html, January 29, 2025.

[39]: Joe Wilkins, (2025), "Team Says They've Recreated DeepSeek's OpenAI Killer for Literally $30. "The results: it just works!"", Futurism, https://futurism.com/researchers-deepseek-even-cheaper, January 30, 2025.

[40}: Kyle Orland, (2025), "OpenAI hits back at DeepSeek with o3-mini reasoning model. OpenAI says faster, more accurate STEM-focused model will be free to all users.", ArsTechnica, https://arstechnica.com/ai/2025/01/openai-hits-back-at-deepseek-with-o3-mini-reasoning-model/, January 31, 2025.

[41]: Dylan Patel, AJ Kourabi, Doug O'Laughlin and Reyk Knuhtsen, (2025), "The DeepSeek Narrative Takes the World by Storm", semianalysis, https://semianalysis.com/2025/01/31/deepseek-debates/, January 31, 2025.

[42]: Trevor Jennewine, (2025), "Nvidia Stock Investors Just Got Good News From President Donald Trump and Wall Street", MSN, Nvidia Stock Investors Just Got Good News From President Donald Trump and Wall Street. Retrieved on January 30, 2025.

[43]: Stephane H. Maes, (2022), "CO2 and CH4 absorption powered by nuclear fusion, via fission, is the only way to manage climate change and the Planet's trigger points", , https://shmaes.wordpress.com/2022/04/09/co2-and-ch4-absorption-powered-fission-is-the-only-way-to-manage-climate-change-and-the-planets-trigger-points/, April 9, 2022, https://osf.io/5ymds.

[44]: Laura Paddison, (2024), "ChatGPT's boss claims nuclear fusion is the answer to AI's soaring energy needs. Not so fast, experts say", CNN, https://www.cnn.com/2024/03/26/climate/ai-energy-nuclear-fusion-climate-intl/index.html, March 26, 2024.

[45]: Justine Calma, (2025), "AI is 'an energy hog,' but DeepSeek could change that. DeepSeek claims to use far less energy than its competitors, but there are still big questions about what that means for the environment", The

Verge, https://www.theverge.com/climate-change/603622/deepseek-ai-environment-energy-climate, January 31, 2025.

[46]: Stephane H. Maes, (2024), "The environmental cost of GenAI is out of control. It triples emissions of data centers!", , September 10, 2024, and other comments on [43].

[47]: Angela Yang, (2025), "On DeepSeek, you can watch AI navigate censorship in real time. A Chinese Embassy spokesperson said, "Artificial intelligence is not outside the law, and all governments are managing it according to law, and China is no exception.", NBC New, https://www.nbcnews.com/tech/innovation/deepseek-censorship-china-rcna189594, January 31, 2025.

[48]: PC World, (2025), "ChatGPT's advanced AI costs $200/mo. Now it's free for Windows users", MSN, https://www.msn.com/en-us/news/technology/chatgpt-s-advanced-ai-costs-200-mo-now-it-s-free-for-windows-users/ar-AA1y8MWK. Retrieved on January 30, 2025.

[49]: Chris Smith, (2025), "DeepSeek AI collects tons of data about you and sends it all to China", BGR, https://bgr.com/tech/deepseek-ai-collects-tons-of-data-about-you-and-sends-it-all-to-china/, January 28, 2025.

[50]: Andrew Thompson, (2025), "Why did DeepSeek tell me it's made by Microsoft? The Chinese-language model has shocked and awed the American stock market. But my chat with it indicates there are many reasons to be skeptical.", FastCompany, https://www.fastcompany.com/91267647/deepseek-told-me-made-by-microsoft-r1-openai-claude-anthropic-ai-model-copilot, January 28, 2025.

[51]: Reuters, (2025), "Microsoft probes if DeepSeek-linked group improperly obtained OpenAI data, Bloomberg News reports", https://www.reuters.com/technology/microsoft-probing-if-deepseek-linked-group-improperly-obtained-openai-data-2025-01-29/, January 29, 2025.

[52]: Emmet Lyons, (2025), "DeepSeek AI raises national security concerns, U.S. officials say", CBS News, https://www.cbsnews.com/news/deepseek-ai-raises-national-security-concerns-trump, January 29, 2025.

[53]: AP, "Did DeepSeek copy ChatGPT to make new AI chatbot? Trump adviser thinks so", https://apnews.com/article/deepseek-ai-chatgpt-openai-copyright-a94168f3b8caa51623ce1b75. Retrieved on January 29, 2025. Now also at https://www.newsbreak.com/the-associated-press-510077/3782434624981-did-deepseek-copy-chatgpt-trump-adviser-thinks-so.

[54]: Laura Italiano and Natalie Musumeci, (2025), "OpenAI has little legal recourse against DeepSeek, tech law experts say", Business Insider, https://www.businessinsider.com/openai-little-legal-recourse-against-deepseek-tech-law-experts-2025-1. Retrieved on February 1, 2025.

[55]: Dario Amodei, (2025), "On DeepSeek and Export Controls", https://darioamodei.com/on-deepseek-and-export-controls, January 2025. Retrieved on January 31, 2025.

[56]: Ben Wodecki, (2025), "DeepSeek failed all safety tests, responding to harmful prompts, Cisco data reveals", Capacity, https://www.capacitymedia.com/article/deepseek-failed-all-safety-tests-responding-to-harmful-prompts-cisco, February3, 2025.

[57]: Tor Constantino, (2025), "4 Warnings About DeepSeek You Need To Know Before Using It", Forbes, https://www.forbes.com/sites/torconstantino/2025/02/04/4-warnings-about-deepseek-you-need-to-know-before-using-it/, February 4, 2025.

[58]: /r/CloudAI, (2025), " Deepseek is heavily overrated IMO, give me your opinion. Sonnet still better with API", Reddit, https://www.reddit.com/r/ClaudeAI/comments/1iaad06/deepseek_is_heavily_overrated_imo_give_me_your/. Retrieved on February 7, 2025.

[59]: Bloomberg, (2025), “The DeepSeek AI revolution has a security problem”, CISO – The Economic Times, https://ciso.economictimes.indiatimes.com/news/cybercrime-fraud/the-deepseek-ai-revolution-has-a-security-problem/117957738, February 6, 2025.

[60]: /r.SillyTarvernAi, (2025), “Is DeepSeek R1 largely unusable for the past week or so? Or does it simply dislike me?”, Reddit, https://www.reddit.com/r/SillyTavernAI/comments/1iirpc3/is_deepseek_r1_largely_unusable_for_the_past_week/. Retrieved on February 7, 2025.

[61]: Will McCurdy, (2025), “DeepSeek Fails Researchers' Safety Tests. 'DeepSeek R1 exhibited a 100% attack success rate, meaning it failed to block a single harmful prompt,' Cisco says.”, PCMag, https://www.pcmag.com/news/deepseek-fails-every-safety-test-thrown-at-it-by-researchers, February 1, 2025.

[62]: Paul Kassianik, Amin Karbasi. (2025), “Evaluating Security Risk in DeepSeek and Other Frontier Reasoning Models”, Cisco, https://blogs.cisco.com/security/evaluating-security-risk-in-deepseek-and-other-frontier-reasoning-models, January 31, 2025.

[63]: Rosie Hoggmascall, (2025), “DeepSeek: when UI doesn’t matter. Localisation, prompting and a cute little whale.”, Medium, https://uxdesign.cc/deepseek-when-ui-doesnt-matter-77dd635c65a6?gi=adb92bf48c45, January 31, 2025.

[64]: Gal Nagli, (2025), “Wiz Research Uncovers Exposed DeepSeek Database Leaking Sensitive Information, Including Chat History. A publicly accessible database belonging to DeepSeek allowed full control over database operations, including the ability to access internal data. The exposure includes over a million lines of log streams with highly sensitive information.”, Wiz, https://www.wiz.io/blog/wiz-research-uncovers-exposed-deepseek-database-leak, January 29, 2025.

[65]: Paulius Grinkevičius, (2025), “The total cost of DeepSeek’s AI models exceeded $1.5 billion, report estimates”, Cybernews, https://cybernews.com/news/cost-of-deepseeks-ai-models/, February 3, 2025.

[66]: Hayden Field, (2025), “DeepSeek's hardware spend could be as high as $500 million, new report estimates”, CNBC, https://www.cnbc.com/2025/01/31/deepseeks-hardware-spend-could-be-as-high-as-500-million-report.html, January 31, 2025.

[67]: Kevin Okemwa, (2025), “"We made a mistake in not being more transparent": OpenAI secretly accessed benchmark data, raising questions about the AI model's supposedly "high scores" — after Sam Altman touted it as "very good". OpenAI's o3 AI model was reportedly trained using a sophisticated benchmark's problems and solutions, potentially explaining its exemplary performance.”, Windows Central, https://www.windowscentral.com/software-apps/we-made-a-mistake-in-not-being-more-transparent-openai-secretly-accessed-benchmark, January 22, 2025.

[68]: David Meyer, (2025), “‘Manipulative and disgraceful’: OpenAI’s critics seize on math benchmarking scandal”, Fortune, https://fortune.com/2025/01/21/eye-on-ai-openai-o3-math-benchmark-frontiermath-epoch-altman-trump-biden/, January 21, 2025.

[69]: McCarthy, J., Minsky, M., C.E. Shannon, (1956), “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence”, http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf.

[70]: E. Brynjolfsson and A. McAfee, (2016), “The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies”, New York: W.W. Norton & Company.

[71]: I. Sutskever, O. Vinyals, and Q. V. Le, (2014), "Sequence to Sequence Learning with Neural Networks," in *Advances in Neural Information Processing Systems* (NIPS), vol. 27, pp. 3104-3112, 2014.

[72]: Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. A., Kaiser, Ł., Polosukhin, I., (2017), “Attention is All You Need”, Advances in Neural Information Processing Systems (NIPS).

[73]: OpenAI, (2023), “GPT-4 Technical Report,” OpenAIhttps://openai.com/index/gpt-4-research/, March 14, 2023 retrived on February 5, 2025.

[74]: Jon Chun, Christian Schroeder de Witt, Katherine Elkins, (2024), “Comparative Global AI Regulation: Policy Perspectives from the EU, China, and the US”, arXiv:2410.21279v1.

[75]: E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, (2021)," On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?, "in Proc. 2021 ACM Conf. Fairness, Accountability, and Transparency (FAccT), Mar. 2021, pp. 610-623.

[76]: Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shinn, J., et al., (2020), “Language Models are Few-Shot Learners”, Advances in Neural Information Processing Systems (NIPS).

[77]: David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, Demis Hassabis, (2017), “Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm”. arXiv:1712.01815v1.

[78]: Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, Dan Mané, (2016), “Concrete Problems in AI Safety”, arXiv:1606.06565v2.

[79]: Nate Nelson, (2025), “DeepSeek Jailbreak Reveals Its Entire System Prompt. Now we know exactly how DeepSeek was designed to work, and we may even have a clue toward its highly publicized scandal with OpenAI. “, Dark Reading, https://www.darkreading.com/application-security/deepseek-jailbreak-system-prompt, January 31, 2025.

[80]: Perplexity (2025), “Perplexity PRO”, https://perplexity.ai. Retrieved on January 1, 2025.

[81]: Patricio Cerda Mardini, Martyna Slawinska, (2025), "Navigating the LLM Landscape: A Comparative Analysis of Leading Large Language Models," MindsDB, https://mindsdb.com/blog/navigating-the-llm-landscape-a-comparative-analysis-of-leading-large-language-models, January 14, 2025.

[82]: [83]: Jan Lasek, Onur Yilmaz, Chenjie Luo and Chenhan Yu, (2024), "Post-Training Quantization of LLMs with NVIDIA NeMo and NVIDIA TensorRT Model Optimizer," developer.nvidia.com, https://developer.nvidia.com/blog/post-training-quantization-of-llms-with-nvidia-nemo-and-nvidia-tensorrt-model-optimizer/, Sept. 10, 2024.

[84]: Kartik Talamadupula, (2024), “A Guide to Quantization in LLMs”, Symbl.ai, https://symbl.ai/developers/blog/a-guide-to-quantization-in-llms/, February 21, 2024.

[85]: Nisha Arya, (2022), “Decision Tree Pruning: The Hows and Whys. Decision trees are a machine learning algorithm that is susceptible to overfitting. One of the techniques you can use to reduce overfitting in decision trees is pruning.”, KDnuggets, https://www.kdnuggets.com/2022/09/decision-tree-pruning-hows-whys.html, September 2, 2022.

[86]: Lawrence Rabiner, Biing-Hwang Juang, (1993), “Fundamentals of Speech Recognition”, Peasron College Div.

[87]: Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, John Guttag, (2020), “What is the State of Neural Network Pruning?”, arXiv:2003.03033v1.

[88]: Hongrong Cheng, Miao Zhang, Javen Qinfeng Shi, (2023-2024), “A Survey on Deep Neural Network Pruning-Taxonomy, Comparison, Analysis, and Recommendations”, arXiv:2308.06767v2.

[89]: Tiernan Ray, (2025), “Apple researchers reveal the secret sauce behind DeepSeek AI. The AI model that shook the world is part of a broad trend to squeeze more out of chips using what's called sparsity”, ZSNet, https://www.zdnet.com/article/apple-researchers-reveal-the-secret-sauce-behind-deepseek-ai/, January 28, 2025.

[90]: Samira Abnar, Harshay Shah, Dan Busbridge, Alaaeldin Mohamed Elnouby Ali, Josh Susskind, Vimal Thilak, (2025), “Parameters vs FLOPs: Scaling Laws for Optimal Sparsity for Mixture-of-Experts Language Models”, arXiv:2501.12370v2.

[91]: Stephane H. Maes, (2024), “MultiAI Document Generation Optimized for task and context”, Internal Report, Factwise.ai, August 12, 2024.

[92]: Stephane H. Maes, (2024), ” MultiAI Document Generation Optimized for task and context at Intelligine.ai”, Intelligine.ai, October 1, 2024.

[93]: Stephane H. Maes, (2024), “From MultiAI to SingleAI honed to a specific domain, user/enterprise, and/or specialized for a specific type of document or media”, Intelligine.ai, November 2024.

[94]: Stephane H. Maes, (2024), “Fixing Reference Hallucinations of LLMs”, https://doi.org/10.5281/zenodo.14543939, https://shmaes.wordpress.com/2024/11/29/fixing-reference-hallucinations-of-llms/, November 29, 2024. (osf.io/u38w4/, viXra:2412.0149v1).

[95]: Multimodal, (2024), “Open-Source AI vs. Closed-Source AI: What’s the Difference? Can’t decide between open-source AI vs. closed-source AI? Learn the key differences and make the best choice for your business.”, https://www.multimodal.dev/post/open-source-ai-vs-closed-source-ai, August 8, 2024.

[96]: Fiona McDonnell, (2025), “What is open-source AI?”, https://telnyx.com/resources/what-is-open-source-ai, January 27, 2025.

[97]: Bill Doerrfeld, (2024), “Be careful with ‘open source’ AI. Open source AI models may be appealing for developers, but there are still plenty of complex risks to assess.”, https://leaddev.com/technical-direction/be-careful-open-source-ai, August 20, 2024.

[98]: Alistair King, (2024), “Op-ed: The Benefits Of Open Source AI”, The Fieldson News, https://fieldstonnews.com/home/2024/05/op-ed-the-benefits-of-open-source-ai/, May 10, 2024.

[99]: Oğuz Kağan Aydın, (2024), “Open Source AI vs. Proprietary AI: Pros and Cons for Developers”, https://www.novusasi.com/blog/open-source-ai-vs-proprietary-ai-pros-and-cons-for-developers, August 13, 2024.

[100]: MoesiF, (2024), “Benefits and Applications of Open Source AI”, https://www.moesif.com/blog/technical/api-development/Open-Source-AI/, July 29, 2024.

[101]: IBM-2024, (2024), “How Open-Source AI Drives Responsible Innovation”, The Atlantic, https://www.theatlantic.com/sponsored/ibm-2024/how-open-source-ai-drives-responsible-innovation/3894/. Retrieved on February 8, 2025.

[102]: OpenAI, (2023-204), "GPT-4 Technical Report", arXiv:2303.08774v5.

[103]: B. Veldman, (2024), "Deploy Azure OpenAI & LLM via Azure Bicep", Medium, https://cloudtips.nl/deploy-azure-openai-llm-via-azure-bicep-2c142bcff7d8, August 1, 2024.

[104]: CSET, (2024), "Key Concepts in AI Safety: Reliable Uncertainty Quantification in Machine Learning,” Center for Security and Emerging Technology (CSET), https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-reliable-uncertainty-quantification-in-machine-learning/, June 2024.

[105]: Gallifant J, Fiske A, Levites Strekalova YA, Osorio-Valencia J S, Parke R, Mwavu R, Martinez N, Gichoya JW, Ghassemi M, Demner-Fushman D, McCoy LG, Celi LA, Pierce R., (2024), “Peer review of GPT-4 technical report and systems card”, PLOS Digit Health. 2024 Jan 18; 3 (1).

[106]: Microsoft, "What's new in Azure OpenAI Service?", Microsoft Learn, https://learn.microsoft.com/en-us/azure/ai-services/openai/whats-new, Jan. 30, 2025.

[107]: Microsoft, "Azure OpenAI Service models", Microsoft Learn, https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models, Jan. 30, 2025.

[108]: T. B. Brown et al., (2020), "Language Models are Few-Shot Learners,", arXiv:2005.14165v4.

[109]: Stephane H. Maes, (2024), "Systems and Methods for Orchestrating Multiple Artificial Intelligence Engines," Intelligine.ai report, Dec. 2024.

[110]: Anthropic Product, (2024), “Building effective agents”, Anthropic, https://www.anthropic.com/research/building-effective-agents, December 19, 2024.

[111]: R. Bommasani et al., (2021-2022), "On the opportunities and risks of foundation models", arXiv:2108.07258v3.

[112]: G. Marcus, (2020), "The next decade in AI: Four steps towards robust artificial intelligence", arXiv:2002.06177v3.

[113]: Amr Elmeleegy, Lequn Chen and Kevin Hu, (2024), "Spotlight: Perplexity AI serves 400 million search queries a month using NVIDIA inference stack", NVIDIA Developer Blog, https://developer.nvidia.com/blog/spotlight-perplexity-ai-serves-400-million-search-queries-a-month-using-nvidia-inference-stack/, December 5, 2024.

[114]: StackOverFlow Blog, (2021), “Open source has a funding problem. Relying on volunteers to maintain every open source project isn't long term sustainable. Funding open source projects could keep development moving, but would that funding be raised and who would pay for it?”, StackOverflow, https://stackoverflow.blog/2021/01/07/open-source-has-a-funding-problem/, January 7, 2021.