SIA

Analysis of trends and strategic perspectives for companies investing in AI

December 2025

/ Executive Summary

- AI and Generative AI (GenAI) have undergone a rapid transformation, from being an academic research breakthrough to being a critical business imperative. This white paper presents a comprehensive analysis of recent developments in the field, providing people and organizations with a strategic framework for harnessing AI's transformative capabilities while cutting through the prevalent hype and deliver actionable intelligence.
- Recent advancements have significantly enhanced AI system capabilities, particularly reasoning abilities and cost-effectiveness, making sophisticated AI tools more accessible to organizations of all sizes. The AI & Tech industry has pivoted toward optimizing model efficiency and developing more compact architecture. While there have been improvements, ongoing challenges like data quality and scarcity still demand innovative solutions, such as synthetic data generation.
- Building on these improvements, Retrieval-Augmented Generation (RAG) has emerged as the predominant implementation of GenAI in companies, enabling large models to supplement their general knowledge with private and confidential sources of information.
- Agentic workflows now represent the next evolution of AI systems, enabling AI models to interact with their environment and business context by using relevant tools. This facilitates more advanced solutions in field where human expertise remains essential. Their development will refocus interest on smaller models, which are faster, more effective and more energy efficient for specific tasks.
- The AI & Tech sector shows remarkable investor excitement and revenue growth that exceeds previous software-as-a-service business patterns, though questions persist about valuation sustainability and long-term profitability. As this market expands, the heterogeneous evolution of regulatory frameworks across jurisdictions creates significant compliance challenges for organizations implementing AI solutions globally. Despite these regulatory complexities, industry priorities have clearly shifted from AI foundation model development to creating practical, market-ready AI solutions addressing specific business needs.

In parallel, on-device AI continues gaining traction by offering enhanced privacy and security benefits, making it more valuable for some industries than conventional applications that rely on cloud inference APIs. This orientation extends to user experience, where text-based interfaces have dramatically lowered adoption barriers compared to complex SaaS implementations requiring extensive development and integration efforts.

/ GenAI will enable organizations to automate time consuming intellectual processes, improve decision-making, and achieve unprecedented operational efficiencies. This highly competitive and rapidly changing market has enabled various actors to emerge from all regions around the world, ranging from proprietary solutions to open-source initiatives, with increasing emphasis on product integration capabilities. As the technology matures, GenAI is transforming both technical and non-technical roles within organizations - creating specialized positions like AI Engineers or AI Ambassador. At the same time, it is democratizing advanced capabilities for non-technical professionals through intuitive tools that boost efficiency and allow them to shift focus from day-to-day operations to more strategic responsibilities.

Ready to engage in the AI transformation of your business and turn your data into a strategic asset? At Sia AI, we support organizations at every stage of their AI journey: from crafting a clear AI adoption strategy to developing impactful use cases and building capable technical teams. We deliver tailored methodologies and ready-to-use AI solutions to help you navigate this transformative landscape, mitigate risks, and achieve measurable results.

Why this White Paper and who is it for?

The last two to three years have redefined what organizations can achieve with AI. Yet, for every technological leap, there is an equal need for interpretation, prioritization, and practical integration. This white paper was written to bridge that gap and **provide insights that connect this disrupting technology with concrete business outcomes.**

This report is intended for AI & Technology-driven leaders. Data, AI and Engineering executives who are seeking to better understand the implications of recent breakthroughs for their business models, operations, and teams. **It is equally designed for AI-acculturated business leaders** — from Digital Innovation officers to transformation executives — who are integrating the next generation of data and AI-enabled tools in their organizations.

At Sia AI, we build, we experiment, and we learn alongside our clients. Every day we observe AI's impact across all sectors, even as understanding often lags behind technical progress. Between oversimplified narratives and technical debates, decision-makers often lack a grounded perspective to act with confidence. This white paper **cuts through the AI hype** to deliver **insights, figures, and frameworks** that are both technically solid and strategically actionable.

Each of the six sections provides:

- **Deep dives into key AI concepts,** from model optimization to agentic workflows.
- **Business applications and case studies,** illustrating how organizations are already capturing value from these technologies.
- **Recognized metrics and data points** from the AI research and practitioner community, ensuring an evidence-based understanding rather than speculation.

What You Won’t Find in This White Paper

This report is neither a marketing brochure nor a speculative essay on AI’s future. You won’t find futuristic promises, visionary storytelling or lists of “top trends”. You also won’t find any new survey, or forecast about AI and its impact on jobs or industries those are already covered by many reputable sources. Our ambition is different: to clarify, not to predict. We combine the best available data with real-world use cases to give a realistic, experience-based perspective of what AI means for your business today.

You also won’t find **code tutorials, technical implementation guides, or vendor comparisons**. While we discuss architecture, frameworks, and technologies, our purpose is not to teach how to build AI systems. We focus on what building (or integrating) them means for your business.

Finally, this white paper avoids **simplistic narratives** that label AI either a miracle or a menace. Instead, we focus on the nuanced middle ground: how organizations can leverage technological breakthroughs responsibly, profitably, and sustainably — while acknowledging the challenges that come with it.

Message from our Leaders

Professional skills now have a lifespan of two years instead of thirty. Thibaut Guilluy, Director General of France Travail*, shared this observation on Sia's podcast on Hypertransformation. This sentence crystallizes a reality we witness daily: we have moved beyond digital transformation into an era of hypertransformation. AI is changing our relationship to work, accelerating and, in some cases, reshaping entire workstreams.

This acceleration demands a new playbook. At Sia, we have been helping organizations navigate complex changes for over 25 years now, and we recognize 2025 as a defining moment where AI, regulatory evolution, and business model disruption converge and redefine our ways of working, whatever the sector.

We identify three paradigm shifts that are reshaping businesses' competitive advantage:

- **Autonomous AI agents managing end-to-end business processes**
- **Real-time edge AI enabling instant decision-making at scale**
- **Fully Integrated AI governance transforming from a top-down compliance burden to a strategic differentiator**

Our recent conversations with transformation leaders — from integration of AI ethics at MAIF to AI-driven product innovation at Groupe Bel — revealed a common thread: successful AI adoption isn't about technology alone, it's about orchestrating machine capability with human decision-making.

New AI transformation leaders don't just deploy AI solutions; they redesign their companies' entire operating model around it. They understand that technical implementation is only a lever for cultural transformation, advancing at an unprecedented pace.

* French national employment agency that helps people find jobs and supports employers with recruitment.

At Sia AI, we have spent over a decade translating AI's promise into measurable performance. Our multidisciplinary teams—combining business strategists, data scientists, AI and Engineering specialists — bring both European excellence in AI governance and global perspectives on technical innovation. We don't just implement AI strategies; we architect AI solutions that align with regulatory requirements, operational realities, and strategic ambitions.

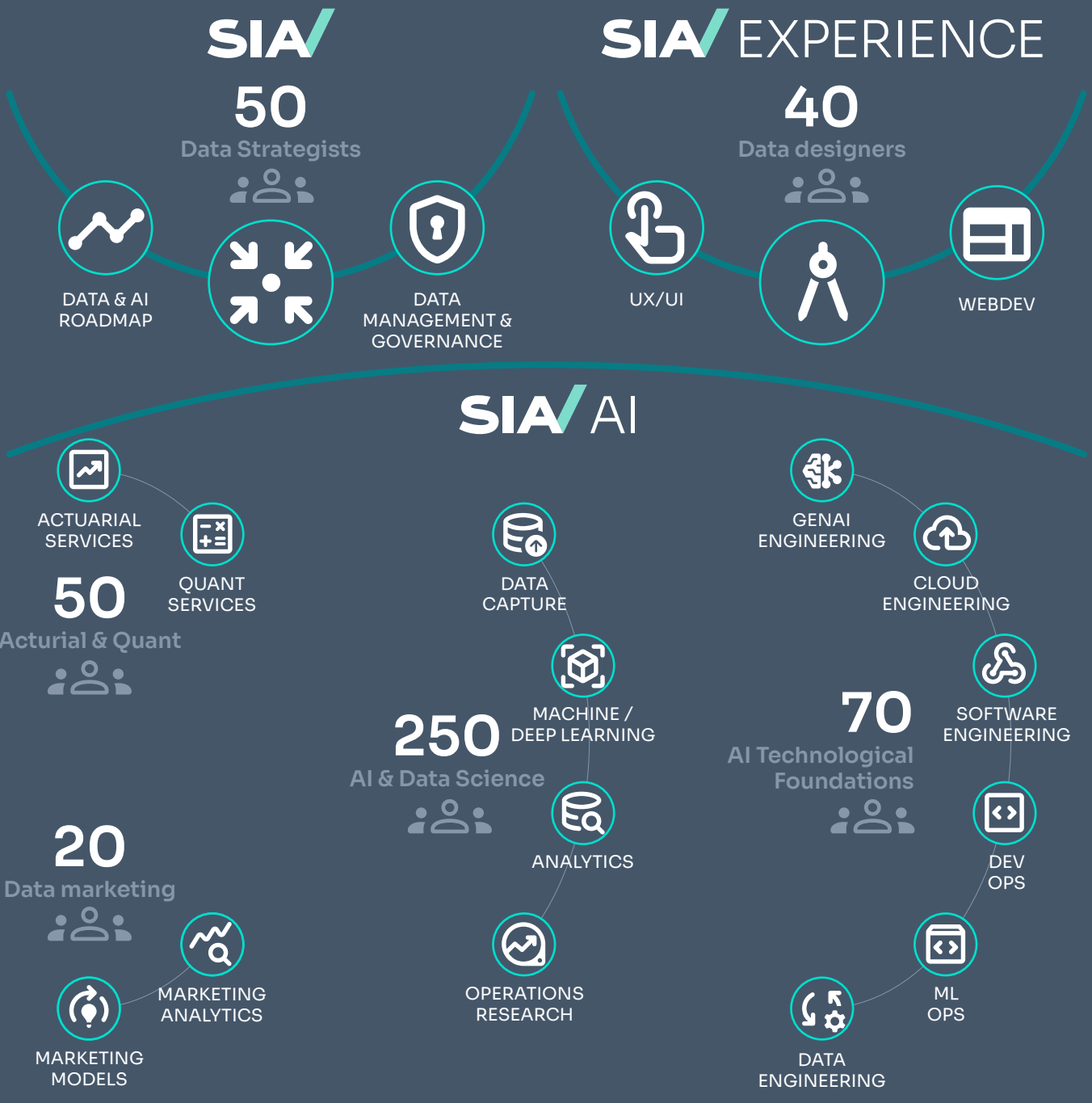
The path forward is clear, organizations that don't just chase trends but shape their own businesses' AI transformation will define the future.

The insights in this white paper represent our collective wisdom from the frontlines of AI cutting edge technologies at the service of business transformation.

Pierre LEPLATOIS, Adrien GRIMAL, Sebastien GERBER, Arnaud TATIN

"At Sia, we are 'Optimists for change'. Let's shape tomorrow, together."

Senior Leadership at Sia AI's Data and Tech Foundation Business Units



/ Key Takeaways

01 Corporate Usage and Change Management:

Organizations that fail to effectively adopt and leverage AI risk missing out on significant productivity gains and will be outpaced by those who do. Developing data literacy and GenAI expertise across all business departments will be essential to achieving successful digital and AI transformation.

02 Keys to AI Adoption and Organizational Transformation:

Effective AI adoption relies on aligning technology with culture, skills, and processes. Organizations that develop their AI expertise, integrate change management, and ensure leadership buy-in are better positioned to translate AI initiatives into tangible business value, while isolated pilot projects are likely to fall behind.

03 Democratized AI Access:

New business opportunities and facilitated access to AI technologies are rapidly emerging, fueled by AI development's focus on efficiency, cost reduction, model size optimization, coupled with the narrowing performance gap between Open-Source and Proprietary Models.

04 Data Quality remains Crucial:

As the volume of training data reaches its limit and the prevalence of synthetic data increases, the quality of training data becomes even more crucial for achieving accurate and reliable AI outcomes. This is also true for AI systems, no matter how effective a model may be; it will never compensate a poor-quality corpus.

05 Multimodal Models Expand AI Capabilities:

Multimodal models, capable of processing and generating various data types (text, images, audio, video), are expanding AI's capabilities and applications across all industries. Organizations will have to rethink their data governance and management policies accordingly.

06 Reasoning is a Key Focus:

AI research is increasingly prioritizing "test time scaling" and the development of "reasoning LLMs" to tackle complex, analytical tasks that require deeper understanding and problem-solving abilities. This will expand AI's applications in the workplace from an advanced data retrieval tool to a genuine work assistant.

07 Agentic Workflows:

Agentic workflows, where AI models use tools and interact with each other or with third-party products, will enable more complex problem-solving and automation. Its large deployment will draw more and more attention to smaller models, which can be faster, cheaper and more efficient for specific tasks.

08 GenAI Growth and Investment:

GenAI companies have experienced explosive user adoption, fueling rapid revenue growth that outpaces traditional SaaS models. This momentum, backed by substantial investment, is accelerating the development of higher-value products and services.

09 Global AI Power Dynamics:

Increasing regulatory scrutiny and geopolitical tensions are reshaping the AI landscape, with China emerging as a major competitor despite hardware constraints, leveraging competitive models and aggressive pricing to establish significant influence amid evolving global governance debates between pro and anti AI regulation.

/ Introduction

Generative AI has undergone a huge transformation in a short amount of time. What was once a research-focused field is now a core business imperative. Recent advancements, driven by breakthroughs in model performance and a significant decrease in costs, have created a critical window of opportunity. Businesses that strategically adopt generative AI now can gain a significant competitive advantage, while those that delay risk losing ground.

Back in 2023, the spotlight was firmly on GPT-4, which dominated benchmarks and captured the imagination of both researchers and businesses alike. Open-source models were behind and struggling to keep pace with their proprietary counterparts. Discussions around AI safety, though important, were just beginning to gain mainstream attention, and political leaders had not yet fully recognized the geopolitical significance of AI.

Fast forward to the present day, the landscape has shifted dramatically. A key development has been the rapid rise of open-source models in terms of performance. Initially focused on catching up to GPT-4, open-source models have now begun to outperform it—all within the span of a year. By 2025, open-source models have taken a clear lead, surpassing proprietary models not only in performance but also in efficiency and size, as demonstrated by models like Gemma-3 12B outperforming several state-of-the-art proprietary models such as Claude 3.7 Sonnet, despite having only 12 billion parameters.

The implications of this shift are profound. It democratizes access to powerful AI capabilities, reduces reliance on single models and software vendors, and opens new possibilities for customization and innovation. This shift towards open-source dominance is not merely a technical detail; it fundamentally alters the strategic implications for businesses when considering AI adoption.

In terms of Research & Development, the community has achieved remarkable breakthroughs at a rapid pace. This includes the development of models that are significantly more versatile, exemplified by the emergence of multimodal systems, and increasingly adaptive, as seen in agentic frameworks that tailor model outputs based on their available tools and environment.

On the industry and business side, GenAI companies, due to their ease of use, have a much higher adoption rate (see **Figure 6**) than traditional SaaS companies, outpacing them in revenue generation (see **Figure 7**). Their significant investments reflect their determination to transform their models into valuable AI products by making them entirely based on GenAI models or by incorporating them into existing solutions. This rapid adoption has led to greater interest from the public and government agencies alike, all of whom are dealing with concerns about ways of working, safety, misuse, intellectual property, and ethics.

In this white paper, we take a step back to look at the full picture behind this tech boom: the R&D and technological leaps, the industry shifts and opportunities, and the evolving political and regulatory responses that have defined recent months. Drawing from these insights, we also outline our predictions for the coming years across these different fields.

The Democratization of GenAI

1. The democratization of GenAI

The era of exclusive access to high-performance AI is giving way to a democratized landscape where businesses can readily adopt and deploy advanced Generative AI solutions. This democratization is characterized by several key trends: the diminishing performance gap between open and proprietary models, the development of essential business use cases like Retrieval-Augmented Generation, and the rise of efficient, compact, and multimodal models. Together, these forces are breaking down traditional barriers to AI adoption, significantly reducing costs and accelerating growth across industries, making sophisticated AI accessible to businesses of all sizes. This section will detail how these innovations are fueling this era of widespread GenAI accessibility and its impact on businesses.

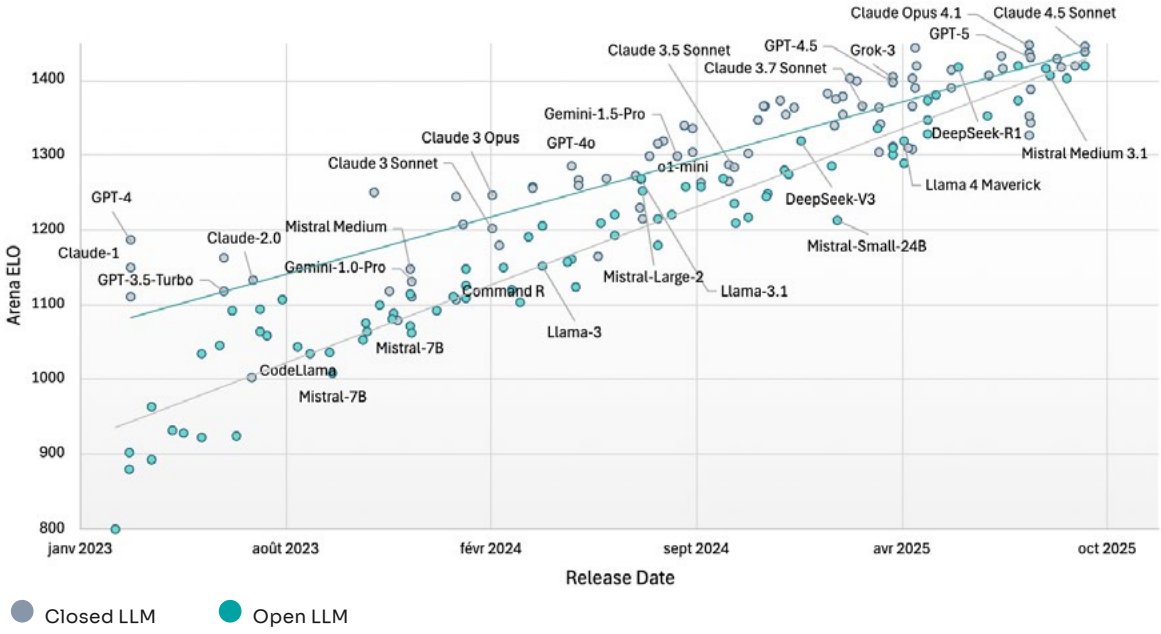
The Performance Gap Between Proprietary And Open-Source Models Decreased

/FIGURE 1: OPEN-WEIGHTS VS PROPRIETARY LLM'S ELO SCORE EVOLUTION BETWEEN 2023 AND 2025

Source: Hugging Face Spaces, Chatbot Arena Leaderboard

Open vs Proprietary LLMs by Arena ELO score

The performance gap between open-source models and proprietary models is decreasing over time, in terms of ELO. ELO is a head-to-head general capability comparison between LLMs (higher is better). It provides a method of comparing LLMs without relying on static benchmarks, enables the capture of fine-grained differences in capability.



Since 2022 and the launch of GPT-3.5, Generative AI has democratized the use of AI among the public and within companies in all sectors. Two main types of models dominate this evolution: "proprietary" models and "open" models. Proprietary models, created by companies with closed and confidential training pipelines, including both the code and training data, prioritize control and exclusivity. In contrast, open-source models[1] provide access to their weights and mostly permissive licenses to deploy them. However, their training data, considered their "secret sauce", is often concealed to maintain their competitive edge. The interaction between these two groups is key to understanding the rapid progress and narrowing of performance gaps in generative AI.

The gap between proprietary and open-source AI models (different from open-weight models[2]) narrowed significantly in 2024. The growing maturity of open-source models makes high-performance AI models more accessible and affordable. Open-source models are closing the gap with proprietary benchmarks due to advancements in knowledge compression, training efficiency, and the use of innovative optimization techniques and larger datasets (see **Figure 1)**.

Meta's release of Llama 3.1's largest version in July 2024 highlighted the rapidly narrowing performance gap between open-source and proprietary models, outperforming other proprietary state-of-the-art (SOTA) models across a range of benchmarks. This progress is particularly evident in mid-range AI models (under 70B parameters), where open-source alternatives are demonstrating increasingly competitive performance.

(1) More rigorously, these are open-weights GenAI models since underlying code, architecture, and training data (fully or partially) are not publicly accessible. We will use this misuse of language to align more closely with the terms commonly used.
(2) Open-source models allow access to everything: the architecture, the training code, the weights and sometimes the dataset. Open-weight models only allow access to the weights or the final trained parameters

Furthermore, open models with fewer than 30B parameters offer impressive inference efficiency and can be deployed on a single high-end GPU. This accessibility provides organizations with a wider range of cost-effective solutions tailored to their specific needs, fostering competition. However, proprietary ultra-large models such as OpenAI's GPT-4 (and more recently GPT-4.5), with their massive parameter counts, continue to excel in ELO benchmarks.

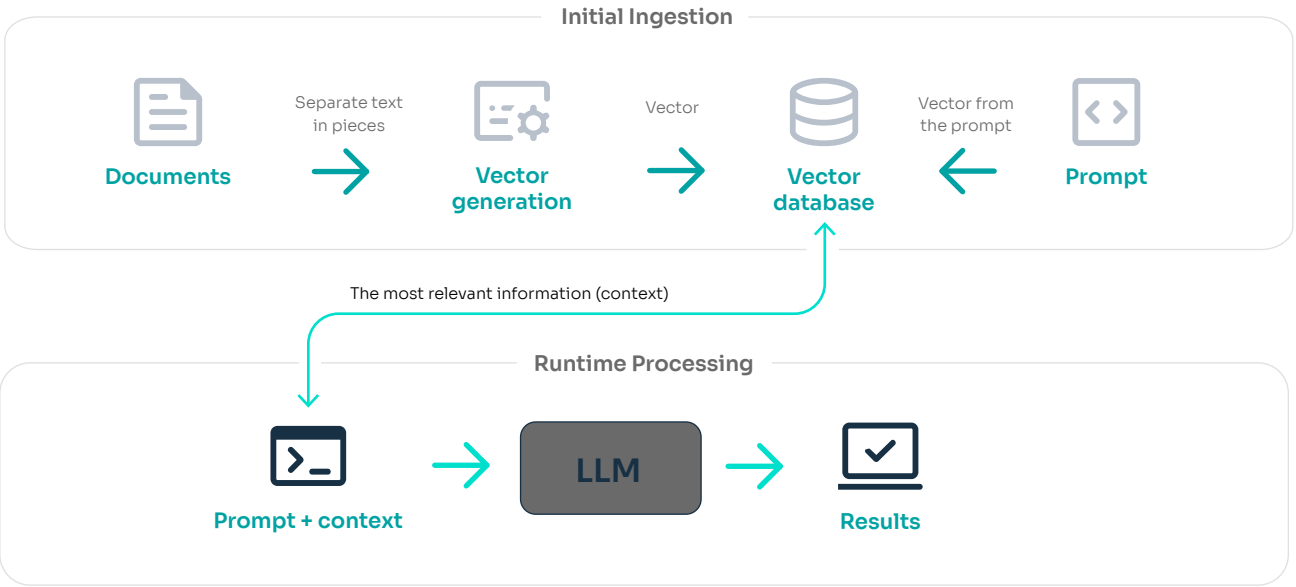
This marks a significant development for businesses, as it democratizes access to high-performance AI. The narrowing performance gap between open-source and proprietary models combined with the growing efficiency of open-source solutions empowers organizations to harness advanced AI capabilities with greater flexibility and still high performance, all at a lower cost. For example, deploying lightweight open models for customer service chatbots can improve response times and user satisfaction, reducing operational costs while enabling tailored solutions for specific industry needs.

Retrieval Augmented Generation (RAG): The Indispensable GenAI Framework Powering Business Use Cases

Early Large Language Models (LLMs) demonstrated impressive general capabilities, but their practical application in business was limited. Organizations struggled to integrate their own proprietary knowledge into these models, creating a significant gap between generic AI capabilities and business-specific value. Retrieval-Augmented Generation (RAG) emerged as a transformative solution, enabling enterprises to leverage the power of LLMs while securely incorporating their unique data assets.

Before RAG's emergence in 2019, incorporating proprietary data into AI systems was a major challenge. Organizations had to rely on resource-intensive methods such as fine-tuning or completely retraining models. These approaches required specialized expertise, significant computational power, and were often prohibitively expensive, preventing many businesses from deploying AI solutions tailored to their unique data.

FIGURE 2: ILLUSTRATION OF A RAG WORKFLOW



The introduction of RAG, further democratized by frameworks like LangChain in 2022, offered a more accessible approach to leveraging LLMs with proprietary data. Instead of relying solely on an LLM's pre-existing knowledge, RAG enables them to access and incorporate information from a company's internal data sources. Critically, while data is transmitted to the LLM's endpoint for processing during inference, RAG enhances privacy by ensuring that this sensitive corporate data is neither stored nor used for subsequent model training. This is achieved by indexing corporate knowledge bases, making them searchable and readily available to the LLM during inference. As a result, the model's responses are grounded in specific, up-to-date, and highly relevant corporate information, with full traceability allowing users to verify the source of each piece of information.

Consider a financial institution seeking to enhance client service, leveraging a RAG-based chatbot, they can deploy a standard open-source model on their proprietary databases and documentation. This system dynamically accesses internal documentation to respond to client inquiries, without requiring retraining of the model on sensitive financial data. This approach maintains strong security: only anonymized text segments are sent to LLM endpoints, with no data retention or training usage. As a result, client interactions are based on up-to-date information, and responses remain traceable to the original source documents, ensuring transparency and accountability.

Recent advancements such as Corrective Retrieval-Augmented Generation (CRAG)[3] further enhance this capability. CRAG introduces verification mechanisms that evaluate and refine retrieved information before integration, significantly improving response accuracy for complex queries.

> RAG has evolved from being merely a proof-of-concept technology to becoming the foundational component of sophisticated enterprise GenAI architecture.”

While extended context windows (see dedicated section below) offer an alternative for certain use cases, organizations can now build comprehensive AI systems where RAG-enabled models remain central to agent-based applications that autonomously navigate complex business processes across vectorized knowledge repositories.

Efficiency and Compact Models: Lowering Costs and Expanding Deployment

The AI model landscape is evolving beyond a simple race for larger, more powerful models. Cost-effectiveness and operational efficiency are now critical factors shaping the competitive dynamics.

> The drive for efficiency is central to the evolution of AI.”

As a result, smaller, more efficient models are gaining significant traction, as demonstrated by HuggingFace's SmolLM, Microsoft's Phi, Arcee AI’s SuperNova-Lite, Google's Gemma, and the Mistral Small family.

(3) Corrective Retrieval Augmented Generation

For businesses, this translates to:

- **Reduced infrastructure complexity and costs:** Smaller models require fewer computing resources compared to larger ones, translating to lower cloud computing expenses and reducing the need for costly on-premise hardware. Because these models can often fit on a single high-performance GPU (in lieu of complex distributed systems), companies can operate with simpler hardware, further reducing implementation and operating costs. At Sia, we believe models in the ~30B parameter range provide a sweet spot between performance and hardware complexity (single GPU) while those in the 70B-100B parameter range provide enhanced performance needed for specific tasks.
- **Edge Computing Capabilities:** AI models can now be deployed directly on devices such as smartphones, IoT devices, or embedded systems. This enables real-time processing without relying on constant cloud connectivity. It improves privacy and reduces latency. For example, Google's Gemma-3 4B, released in mid-2025, demonstrates AI model optimized for on-device computing, outperforming original GPT-4.
- **Better Latency and expanded options:** The increased speed of AI models enhances user experience by reducing response latency. It enables new time-sensitive applications. The development of open-source frameworks like vLLM[4], NVIDIA NIM[5], and the recently introduced SGLang[6], are pushing the boundaries of inference speed, opening up new possibilities for real-time AI deployment. Additionally, frameworks such as ZenML[7] enable structured deployment of LLMs with features like reproducibility and cloud integration. Llama.cpp[8] offers cost-effective, eco-friendly, and flexible compute options.

Simultaneously, techniques such as distillation, quantization, and other model architecture optimizations help create smaller models that are easier to host while maintaining reasonable performance.

On-device AI is on the rise

Building on the edge computing capabilities discussed previously, AI models can now be deployed directly on personal devices. **The GenAI landscape has been seeing remarkable progress, combining higher performance with significantly smaller models**, a trend driven by the growing demand for on-device inference, confidential processing and reduced latency. Indeed, as seen in Figure 3, by making the assumption that cheaper models are generally smaller models, we observe that smaller models tend to outperform larger ones over time in the Elo Arena text score. This capability now extends to smartphones, unlocking secure and private AI applications directly on personal devices.

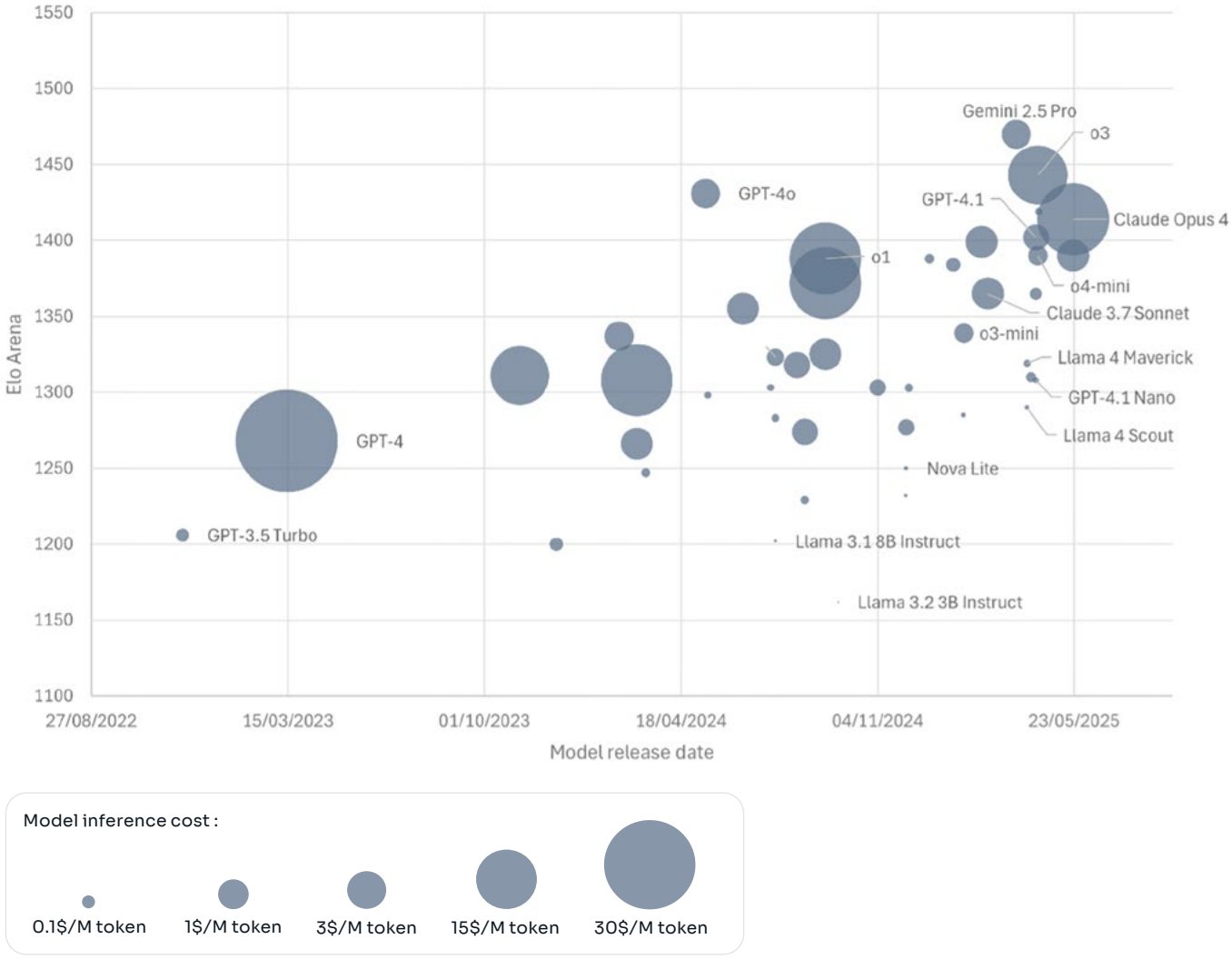
The market has also seen the emergence of other compact but powerful models from leading AI providers, including Hugging Face's SmolLM, Microsoft's Phi, Meta's compact Llama, Mistral's Ministral, and Alibaba's Qwen. Google's Gemma 3 series - particularly the 4B and 1B versions - allows Android developers to integrate real-time AI capabilities directly on-device, eliminating the need for continuous cloud connectivity.

Taking this a step further, Google has introduced the gemma.cpp library, which enables Gemma models to run efficiently on CPUs through optimized inference. The development of such a library highlights the interest of leading AI providers in on-device AI.

The above advances have been enabled by state-of-the-art hardware innovations, including next-generation tensor processing units (TPUs) and GPUs, specially designed to optimize the performance of compact AI models. Ultimately, these improvements make generative AI not only more powerful but also more accessible, enabling it to meet the needs of a wider range of applications and users.

(4) https://docs.vllm.ai/en/latest/
(5) https://developer.nvidia.com/nim
(6) https://docs.sglang.ai/
(7) https://www.zenml.io/
(8) https://github.com/ggml-org/llama.cpp

/FIGURE 3: ELO ARENA SCORE ON RELEASE DATE WITH INFERENCE COST DIMENSION



Multimodal Models: Unlocking New Dimensions of AI Interaction

While LLMs offer powerful text processing and generation capabilities, their impact within corporate ecosystems can be limited. The primary challenge for organizations is integrating and powering applications with AI to drive tangible business value. Multimodal models are a key enabler in this transformation, extending traditional text-based capabilities to process and generate text, images, videos, and audio. This pivotal shift moves beyond text-only AI, creating systems capable of understanding and interacting with users in richer, more human-like ways, even surpassing human capabilities on certain tasks.

Recent advancements underscore the growing importance of multimodal capabilities. OpenAI's SORA and Google's Veo 3, for example, lead the field in generating realistic videos with synchronized sound effects. These innovations are not just technological milestones; they redefine possibilities for sectors like advertising, education, and creative industries. In marketing, text-to-image (TTI) and text-to-video (TTV) capabilities streamline campaign development. Similarly, text-to-speech (TTS) functionalities can enhance the analysis and improvement of customer calls in call centers, support training scenarios, and optimize common scripts to better understand client needs.

The ability to process and generate multiple modalities of data opens a wide range of applications across various industries:

- **Advanced Analytics:** Multimodal models provide richer insights by merging visual and textual data. For example, they can correlate security footage with incident reports to optimize protocols, enhancing risk management.
- **Enhanced Customer Experiences & Accessibility:** Multimodal models can create intuitive customer interactions by understanding both spoken and visual cues. For example, they can generate accessible features such as image descriptions for visually impaired users or real-time transcription in customer service.
- **Automated Content Creation and Streamlined Operations:** By combining different types of input, these models can automate the creation of marketing content or training materials. In industries like construction, they can analyze video feeds to identify hazards, improving safety and efficiency.

Recent advances have both significantly reduced costs and expanded context window. This makes it more feasible to process inputs like PDFs directly and have lessened the dependency on RAG. Nonetheless, RAG remains valuable for enhancing response relevance and grounding outputs in domain-specific knowledge, especially in enterprise applications where precision and context are critical.

By breaking down barriers between input types and incorporating advanced contextualization techniques, multimodal AI offers a pathway to improve corporate workflows, drive innovation, and enhance decision-making across sectors.

The Declining Cost of AI: A New Era for Business

AI models are advancing rapidly, becoming both more powerful and affordable. In June 2023, OpenAI's GPT-4 cost $50 per million tokens[9]. By early 2025, GPT-4o dropped to $10 per million tokens, while competitors like DeepSeek R1 and Gemini 1.5 Flash shattered price barriers at $1 and 12 cents per million tokens, respectively. Although not all these models are equal in performance, each of these newer models outperforms the original GPT-4 from June 2023 (see **Figure 4**).

> This rapid decrease in cost, coupled with increasing performance, is changing the economics of AI adoption."

This shift is driven by advancements in model efficiency, scalability, and a growing focus on frugality. Developers are optimizing software to require less expensive hardware, while increased competition among providers has pushed prices down, benefiting end customers.

Lenovo's implementation of Studio AI exemplifies this trend, demonstrating how generative AI enables organizations to extract maximum value from existing assets. By leveraging their proprietary product libraries and company resources within a structured AI framework, Lenovo reduced annual marketing content creation costs from $42 million to $4.2 million while maintaining quality output tailored for IT decision-makers.

This is not just a question of incremental improvements; it's about unlocking entirely new business possibilities

> AI Use cases that were previously cost-prohibitive are now becoming viable."

(9) A token could be approximated as a word or half a word. It really depends on the tokenizer used by the model. With more recent tokenizers it is better to use half a word as an approximation for a token.

Non-exhaustive advantages of this trend are:

Experiment with AI at a Lower Risk

Pilot projects and proof-of-concept initiatives become significantly more affordable.

Scale AI Solutions More Easily

Deploy AI across a wider range of operations and departments without exorbitant costs.

New Products and Services

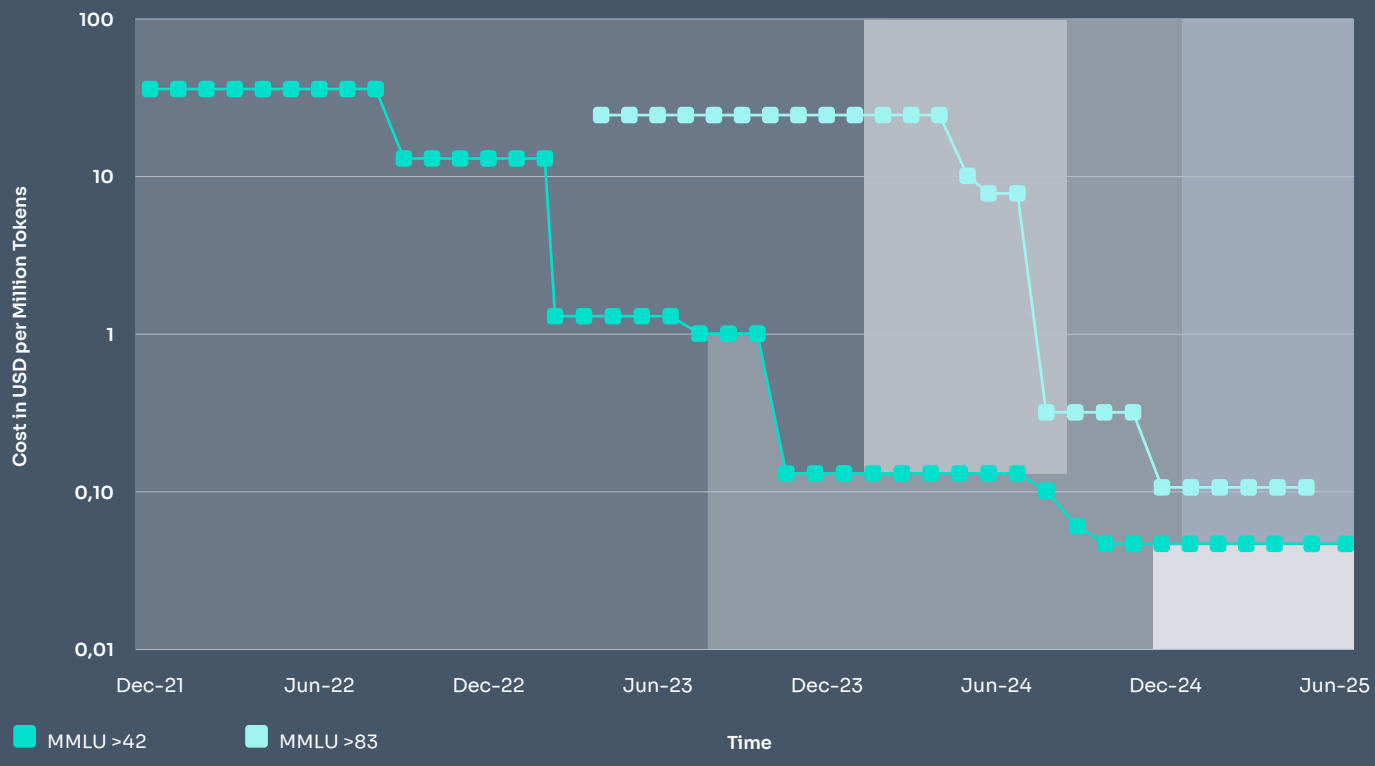
Create innovative offerings that were previously unimaginable due to the high cost of AI processing.

Gain a First-Mover Advantage

Invest now in pilot programs, which can enable a competitive advantage, as prices continue to drop.

FIGURE 4: COST OF THE CHEAPEST LLM WITH A MINIMUM MMLU SCORE (LOGARITHMIC SCALE)

Source: Andreessen Horowitz: LLM inference cost is going down fast



While inference cost per token is decreasing, overall inference costs can still increase in certain situations, such as with test-time scaling for reasoning (see o1 cost in **Figure 5**) and autonomous agents that require multiple iterations (e.g., coding). For instance, online user feedback suggests that Claude Code, a coding agent in Beta Research Preview using Claude 3.7 Sonnet, can be expensive due to the iterative nature of its tasks. Such autonomous agents require numerous iterations to refine their outputs, leading to higher inference costs despite the decreasing per-token rate. Although pricing for this research preview may change, its current cost remains significant due to the demands of its use case. Additionally, OpenAI continues to maintain premium pricing for its flagship models, GPT-4.5 and o1-pro, while reducing prices for models such as GPT-4.1 and o3-mini. The rest of the players are trying to maintain prices low (see **Figure 5**) to compete and gain users.

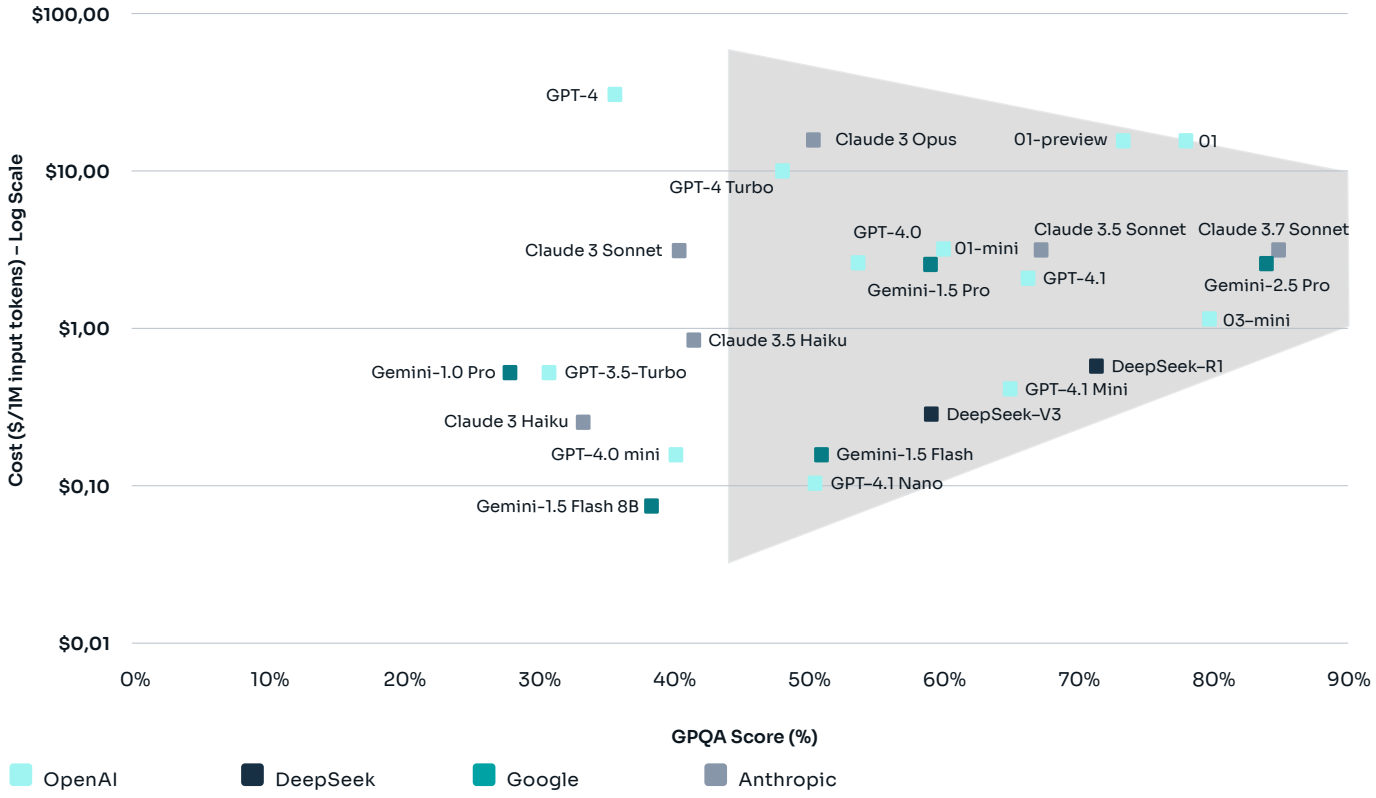
OpenAI might rethink that decision though, as they have decided to discontinue support for the GPT-4.5 API by July 2025 (released towards the end of February 2025), in favor of the cheaper and more recently released GPT-4.1 Series, o4-mini or o3-pro. Time will tell, but the overall trend pushes for better and cheaper models around a standardized cost.

/FIGURE 5: COMPARING LLM COSTS WITH THEIR GPQA SCORES (%)

Source: LLM Stats

Cost vs. Quality comparison between major LLM providers

This graph shows a comparison of prices between different LLM providers vs. their GPQA score. GPQA is a Google-Proof Q&A Benchmark, a challenging benchmark of MCQs in various scientific domains, curated by experts. The questions are high-quality and extremely difficult, hence performance on this benchmark can function as a proxy for LLM.



Businesses should not be deterred by the initial investment costs required to set up knowledge bases and/or the development of an AI-first data culture as the mid-term to long-term potential is clear. This increasingly competitive pricing landscape and downward pricing trend suggest that today's seemingly expensive AI projects are likely to become significantly more cost-effective in the very near future.

From Zero to Scale: The Growth of GenAI in Industries

GenAI is rapidly transforming industries, driven by its ability to lower adoption barriers and enhance efficiency. This is particularly evident in software development, where AI coding assistants are revolutionizing workflows. GitHub Copilot leads the market, while Anthropic's Claude Code and Vercel's V0 have expanded this transformation with in-browser coding and execution capabilities, fundamentally changing how developers interact with AI.

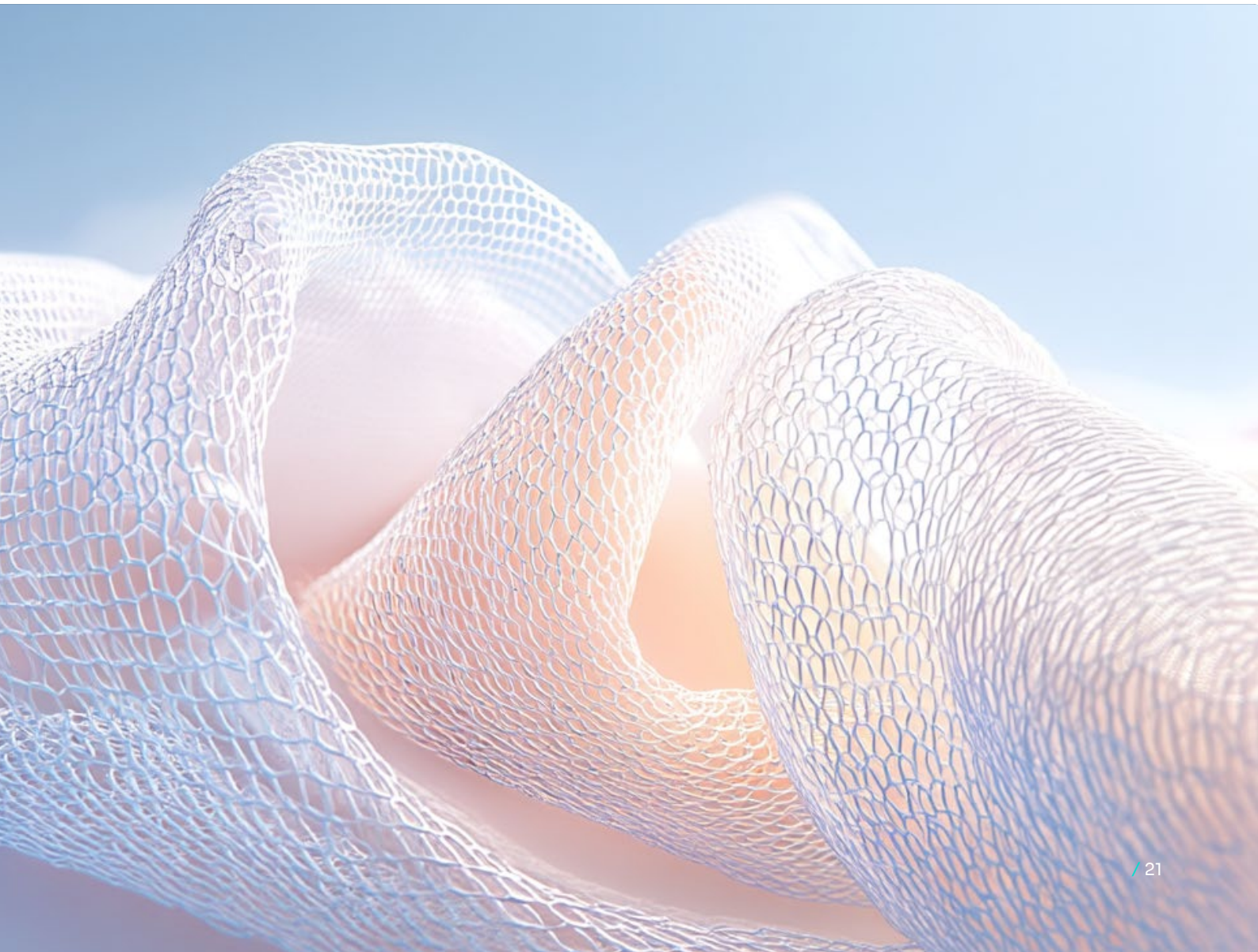
This accelerated growth extends beyond software development. Financial analyses reveal GenAI companies are achieving revenue milestones substantially faster than previous software cycles[10]. According to Stripe data, GenAI companies reach $1 million in annual revenue within 11 months, compared to 15 months for traditional technology companies. These enterprises reach $30 million five times faster than traditional SaaS counterparts, with Cursor scaling from $1 million to $100 million ARR in just one year[11].

This unprecedented growth stems directly from GenAI's intuitive interface: language. By allowing users to access sophisticated capabilities through simple conversational prompts rather than complex UI, organizations implement solutions with minimal training requirements. This accessibility advantage reduces deployment timelines and drives rapid adoption (see **Figure 6**). The technology's ease of use accelerates market penetration and revenue growth (see **Figure 7**). As France's "AI Efficiency" initiative[12], highlighted in February 2025, at least

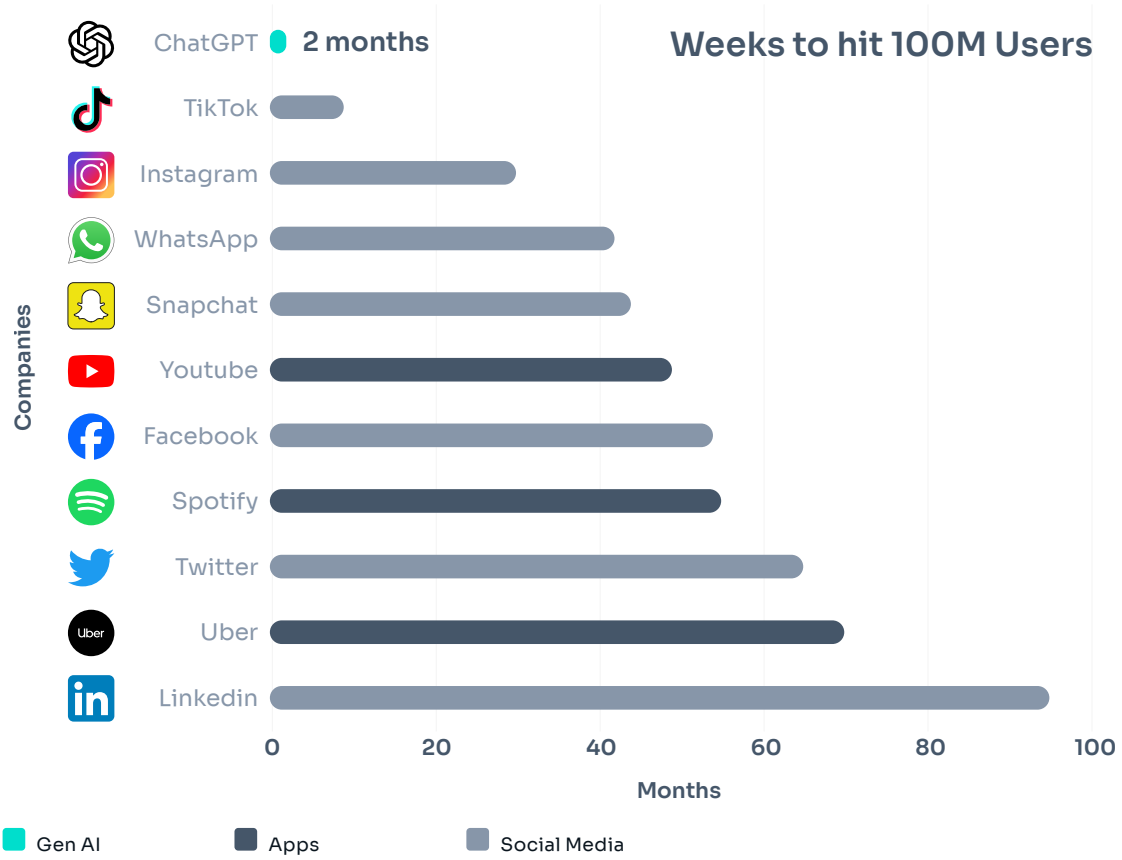
> "55% of submitted projects leverage GenAI technologies, particularly in the services sector (70%)"

These projects frequently incorporate advanced GenAI capabilities such as Retrieval-Augmented Generation (RAG) if applicable, underscoring the strategic adoption of GenAI in operational processes as organizations prioritize transformative technologies to enhance their competitive edge.

(10) Murgia, M. (2024, September 27). AI start-ups generate money faster than past hyped tech companies. Financial Times. Retrieved from https://www.ft.com
(11) Cursor at $100M ARR | Sacra
(12) https://www.entreprises.gouv.fr/la-dge/actualites/ami-ai-efficiency-les-laureats-en-detail



/FIGURE 6: 100M USERS MILESTONE
FOR SOME WELL-KNOWN COMPANIES



/FIGURE 7: REVENUE MILESTONE
FOR SOME WELL-KNOWN GENAI COMPANIES
Source: TechCrunch, Appfigures Explorer

Understanding The Strategic Inflection Point

2. Understanding the Strategic Inflection Point

The Generative AI wave is no longer just about powerful language models; it's about delivering practical and scalable solutions that directly impact business outcomes.
In this section, we dive into the critical advancements enabling this shift, from data challenges like scarcity and quality to pushing the boundaries of model efficiency and reasoning. Agentic AI, coupled with protocols like MCP, is empowering businesses to automate complex workflows, integrate even better with existing systems, and opens new levels of operational efficiency and innovation.

The Shift from AI Models to AI Products, Systems and Services

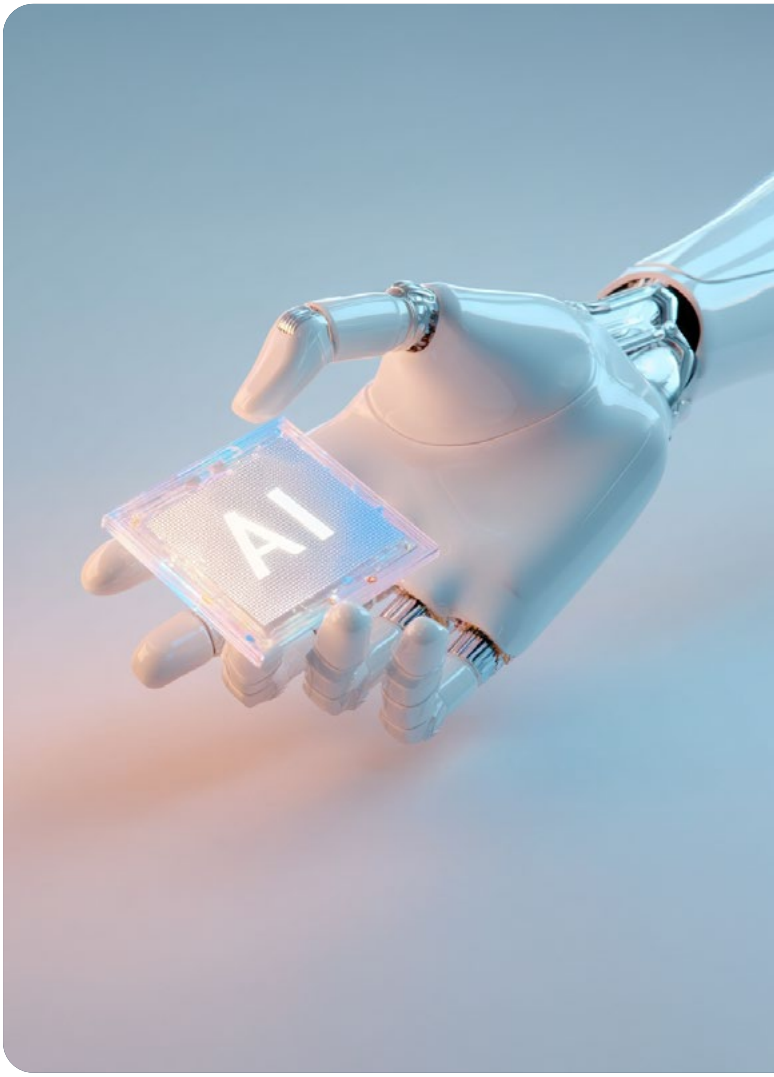
The pursuit of ever more powerful AI models is fueling unprecedented investment in the GenAI space. At the forefront of this, companies such as OpenAI and Anthropic, are committing enormous capital. The scale of this investment is evident in their 2024 financials: OpenAI spent $5 billion while generating $3.7 billion in revenue[13], and Anthropic burned through $5.6 billion, projecting another $3 billion burn in 2025[14]. This massive spending is enabled by significant backing from major investors: Microsoft has invested heavily in OpenAI, who recently raised $40 billion at a $300 billion valuation[15]. Similarly, Amazon and Google have invested in Anthropic, which secured $3.5 billion in additional funding[16].

These organizations then face ever-increasing investor pressure to demonstrate profitability and long-term sustainability. Reflecting this shift, OpenAI also announced plans to transition from a non-profit research organization controlling a for-profit arm to a completely for-profit AI company[17].

Driven by the need for revenue and long-term viability, GenAI companies are shifting their focus from pure research to developing practical, market-ready products such as ChatGPT Enterprise and Microsoft Copilot. In parallel, AI-native companies and Labs like Mistral AI are heavily investing in solutions based on their foundation models. The launch of Le Chat Entreprise in May 2025, a privacy-focused, customizable AI assistant, exemplifies' CEO Arthur Mensch's vision that the next wave of AI chatbots will stand out through innovative features and superior user experience.

With such services, some GenAI-first companies are already seeing strong financial success. ElevenLabs, the leading text-to-speech provider, hit a $1.1 billion valuation in January 2024 achieving unicorn status, with its tools adopted by 62% of Fortune 500 companies. Its Series C raise of $180 million had it tripling its valuation to $3.3 billion as of January 2025[18]. The recent partnership with Spotify to integrate AI-narrated audiobooks highlights how such collaborations help GenAI firms expand into new markets and boost revenue.

To fully realize this transformative potential, however, organizations must now shift their focus towards a critical resource: data. The next frontier in AI lies not just in model innovation, but in strategically addressing the challenges of data scarcity, quality, and accessibility. A new approach to data-centric AI development has become paramount for AI systems as organizations.



(13) https://www.nytimes.com/2024/09/27/technology/openai-chatgpt-investors-funding.html
(14) https://www.reuters.com/technology/anthropic-projects-soaring-growth-345-billion-2027-revenue-information-reports-2025-02-13/
(15) https://www.bloomberg.com/news/articles/2025-03-31/openai-finalizes-40-billion-funding-at-300-billion-valuation
(16) https://www.investopedia.com/ai-startup-anthropic-valued-at-usd61-5b-after-latest-funding-round-11689703
(17) https://www.gzeromedia.com/gzero-ai/openais-nonprofit-days-are-behind-it
(18) https://www.reuters.com/technology/artificial-intelligence/voice-ai-startup-elevenlabs-closes-new-funding-round-33-billion-valuation-2025-01-30/

The growing challenge of data: acquiring high-quality data in an increasingly scarce landscape

Despite the remarkable advancements in Gen AI, the pace of performance gains is slowing, and new challenges are emerging.

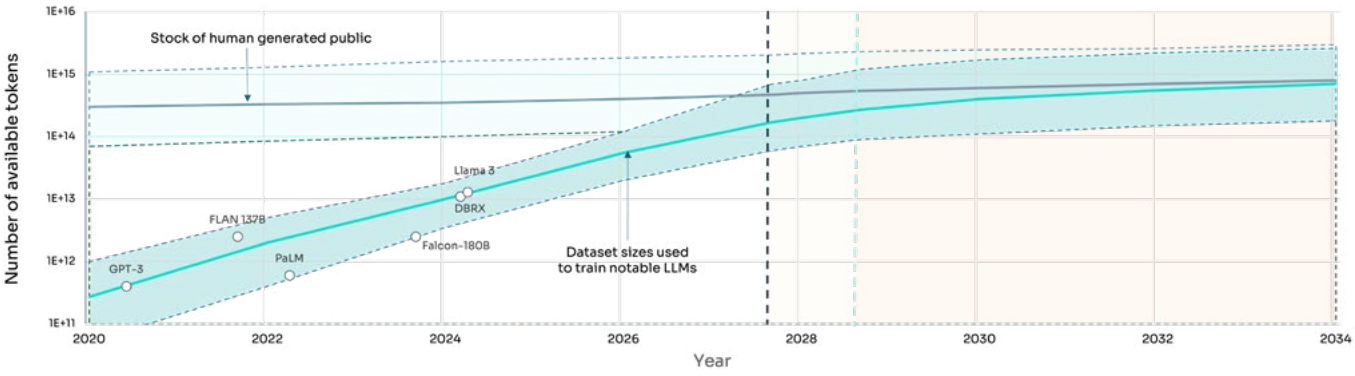
Companies deploying AI tools and applications face these challenges directly, necessitating a strategic shift in how they approach AI development and data reliability.

/FIGURE 8: PROJECTIONS OF THE STOCK OF HUMAN-GENERATED PUBLIC TEXT AND DATASET SIZES USED TO TRAIN NOTABLE LLMS

Source: Will we run out of data? Limits of LLM scaling based on human-generated data, Epoch AI

The Data Scarcity Challenge

Performance improvements driven solely by the volume of data are diminishing as we are running out of human generated data to train on.



- Estimated Stock of Human-Generated Public Text (Tokens)
- Dataset Sizes Used to Train Notable LLMs (Tokens)
- Median date of full stock utilization (5x overtraining)
- Median date of full stock utilization
- Model

Data Scarcity

Performance improvements driven solely by increasing the volume of training data are diminishing. Each new model generation shows smaller performance gains, despite using increasingly larger training datasets. **Figure 8** projects the timeline for AI training data depletion according to both historical and compute-based forecasts[19].

According to this calculation, high-quality language stock data has already run out sometime half-way through 2024. Collecting massive amounts of data is no longer sufficient.

> The competitive advantage will shift to businesses that can access, curate, and train AI models on their unique, high-quality datasets."

/FIGURE 9: PROJECTION OF EXHAUSTION OF AVAILABLE TRAINING DATA

Source: The 2024 AI Index Report

Projections of ML data exhaustion by stock type: median and 90% CI dates

Source: Epoch, 2023 I Table: 2024 AI Index report

Stock type	Historical projection	Compute projection
Low-quality language stock	2032.4 *(years)* [2028.4; 2039.2]	2040.5 *(years)* [2034.6; 2048.9]
High-quality language stock	2024.5 *(years)* [2023.5; 2025.7]	2024.1 *(years)* [2023.2; 2025.3]
Image stock	2046 *(years)* [2037; 2062.8]	2038.8 *(years)* [2032; 2049.8]

(19) The historical projections are based on observed growth rates in the sizes of data used to train foundation models. The compute projections adjust the historical growth rate based on projections of compute availability. Source: 2024 AI Index Report.

Synthetic Data: A double-edged sword for AI

Growing privacy concerns in real-world data collection are driving interest in synthetic data alternatives that mirror the statistical characteristics of authentic datasets while avoiding the associated risks.

> Synthetic data can be a powerful lever to improve model performance."

Consider Microsoft's Phi-4 model, which demonstrates the potential value of curated synthetic data. Phi[20] employs synthetic data as a form of "spoonfeeding," strategically presenting information tailored to the model's current understanding. This new approach suggests that synthetic data, when carefully controlled and deployed, may offer enhanced reasoning capabilities. However, it also exhibited limitations: increased reliance on synthetic data led to underperformance on knowledge-intensive tasks and higher risks of hallucinations.

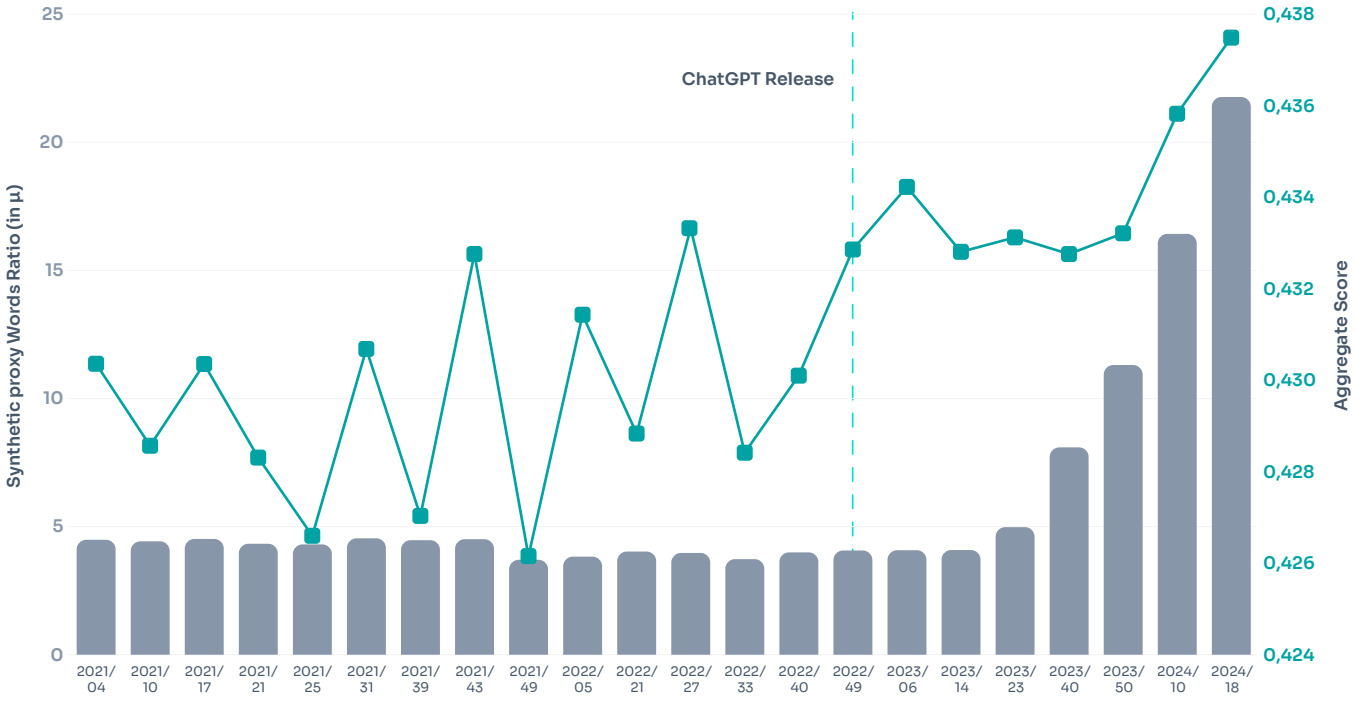
The implications of synthetic data are not uniformly positive. As the internet becomes increasingly populated with AI-generated content, including blog posts and images, the potential for data pollution grows. This synthetic data often lacks the coherence and accuracy of real-world data, potentially degrading the performance of AI models trained on it[21]. This poses a significant risk: models trained on "polluted" data may produce unreliable and biased outputs, leading to flawed decision-making and reputational damage for businesses. Consequently, the effectiveness of relying on recent data alone to improve GenAI performance will likely diminish over time.

FIGURE 10: EVOLUTION OF CONTAMINATED DATA IN COMMON CRAWL DATA

Source: FineWeb by Hugging Face, Open Review

Synthetic Data Contimation: Evolution of contaminated data in Common Crawl over time

Since the release of ChatGPT, there has been a sharp increase in the proxy metric which detects synthetic data (orange), by counting the occurences of words favored by ChatGPT such as "delve", "as a large language model", "rich tapestry", "intertwined", etc.
The score is the evaluation score (0-1, higher is better) of the collected data using the lighteval library (blue). Improved filtration and deduplication techniques are responsible for the increasing evaluation score.



The horizontal axis refers to a version of Common Crawl Data for a particular year.

(20) Phi-4 Technical Report
(21) Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. Nature, 631(8022), 755759.

Overall, we should start decoupling reasoning and intelligence capabilities from pure knowledge compression. While LLMs thrived in the latter thanks to their large training corpus, they show limitations in the former because of their design: statistical prediction of the next token.

Data Quality: The critical differentiator

As the quantity of readily available data diminishes and the prevalence of synthetic data increases, the quality of data collected for training AI models is of paramount importance. Simply put

> "The better the data, the better the AI model, and crucially, the better the products and services."

Curated, high-quality synthetic data offers potential advantages for training LLMs; conversely, synthetic data pollution in public datasets could lead to unreliable outputs. Two key pieces of evidence highlight this:

THE PROBLEM

Contamination of Public Datasets: Widely used public datasets, like Common Crawl[22] are increasingly contaminated with AI-generated content. **Figure 11** illustrates how the proportion of synthetic data in Common Crawl has risen significantly since the release of ChatGPT[23].

THE SOLUTION

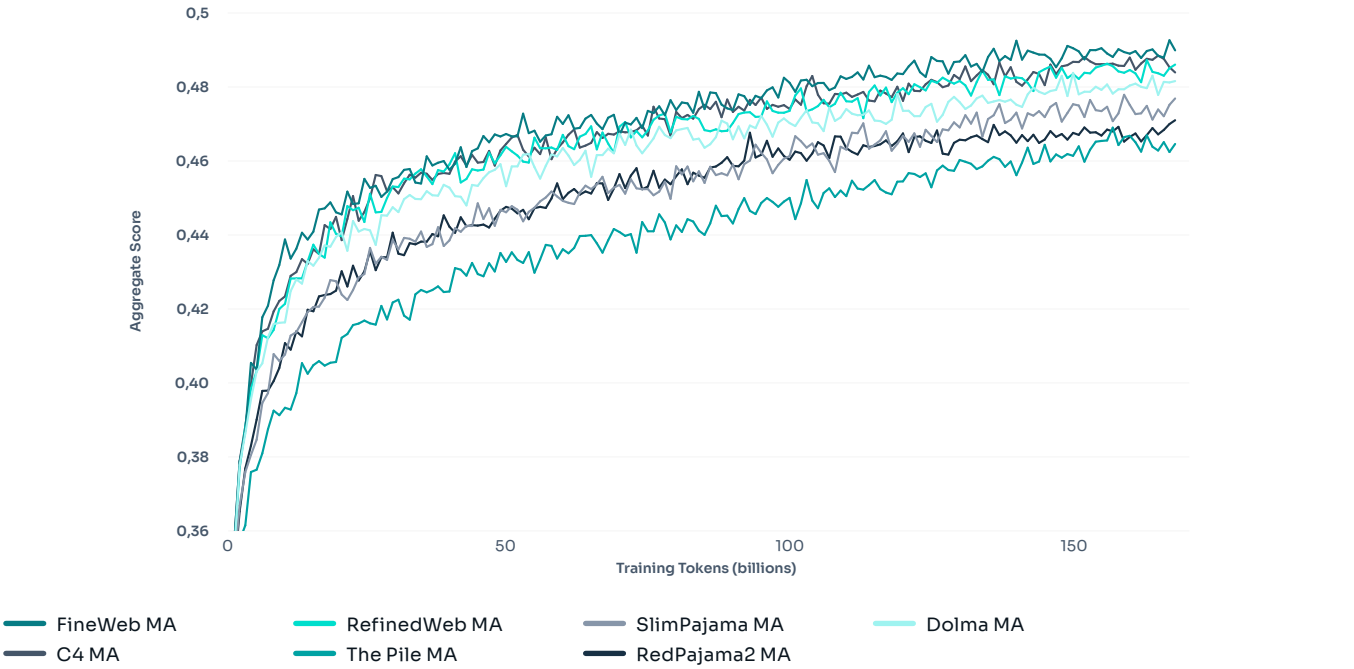
Proven Benefits of High-Quality Data: Research efforts, such as the development of the meticulously curated 15 trillion tokens FineWeb dataset by Hugging Face[24], demonstrate the dramatic impact of data quality on model performance. As illustrated in **Figure 11**, models trained on FineWeb consistently outperform models trained on lower-quality datasets (like C4, The Pile, and RedPajama) across various benchmark tasks. Investing in data quality directly translates to improved AI performance.

FIGURE 11: HIGHER QUALITY DATA LEADS TO BETTER MODEL PERFORMANCE

Source: FineWeb by Hugging Face, Open Review

Higher quality data leads to better model performance

The aggregate score is the aggregated evaluation score of LLMs on various benchmarks, evaluated using the lighteval library.
The legend mentions various internet scale datasets will several billion tokens, that are used to train LLMs. To better compare the datasets, we plot a 3-point moving average (MA) for all the datasets considered.



(22) Common Crawl is a publicly available, massive dataset containing billions of web pages that have been systematically collected and archived from across the internet, widely used for training language models and other AI applications.
(23) FineWeb: Decanting the web for the finest text data at scale - A hugging face space by HuggingFace. (n.d.).
(24) Penedo, G., Kydlíček, H., allal, L. B., Lozhkov, A., Mitchell, M., Raffel, C., ... Wolf, T. (2024b). The FineWeb datasets: Decanting the web for the finest text data at scale.

Although indisputable on the scale of a few billion tokens, this graph does not allow us to conclude that data quality has as much importance in the context of large scale training. However, it clearly shows that a small model trained with little but good quality data can equal or even outpace larger models.

For traditional businesses, the key to effective AI implementation lies not in training from scratch, but in providing LLMs with high-quality context to excel within their specific domain. This requires robust corporate data management focused on collecting, storing, cleaning, preprocessing, validating, and verifying your organization's unique information. Strong data governance is essential for ensuring AI systems can access the right proprietary context, supporting either Retrieval-Augmented Generation (RAG) or targeted fine-tuning with domain-specific data.

> Being able to effectively clean internal knowledge bases is what will ultimately determine AI success in business."

Overall model performance continued to improve

Despite challenges related to data quantity and quality, LLM performance continues to improve and this translates directly to increased business value. Companies can leverage AI to solve more complex problems, automate intellectual tasks, and gain deeper insights from their data.

Automation Solutions

Automate complex, multi-step processes that previously required human judgment.

Personalized Content creation

Deliver highly targeted and relevant marketing messages at scale.

Enhanced Fraud Detection

Identify and prevent fraudulent activities with greater speed and precision.

Regulatory Compliance

Automate compliance checks, reducing risk and freeing up valuable resources.

Advanced information access

Simplify and automate information access in multi-document databases.

While models like GPT-4o showed moderate improvements over their predecessors, other model families and Vision Language Models (VLMs)[25] demonstrated substantial progress. LLMs are now excelling in "reasoning" areas where they were previously limited. For example, in December 2024, OpenAI's o3 model achieved an accuracy of 25% on FrontierMath benchmark[26] – a significant jump from its predecessor's 3%. Same on the ARC-AGI-1 benchmark,[27] performance progressed from 0% in 2023 with GPT-3 to 5% in 2024 with GPT-4o and finally reached 53% by the end of 2024 with o3-medium and 41% by April 2025 with o4-mini-medium[28].

> The progress AI models have made in reasoning ability is a game-changer for businesses, allowing them to tackle complex, multi-stage, intellectual tasks that were previously beyond reach."

This includes advanced data analysis, strategic decision support, and problem-solving that leverages web search, benefiting diverse sectors:

Though OpenAI remains a leader, new and improved model families are emerging, many of which offer distinctive advantages such as lighter architectures and greater energy efficiency.

(25) https://www.vellum.ai/blog/llm-benchmarks-overview-limits-and-model-comparison
(26) Frontier Math evaluates LLM on advanced math reasoning abilities across challenging college and graduate-level problems, requiring multi-step reasoning.
(27) ARC-AGI-1 evaluates advanced reasoning capabilities of AI systems through complex, novel problems requiring adaptability and generalization beyond training data.
(28) Analyzing o3 and o4-mini with ARC-AGI

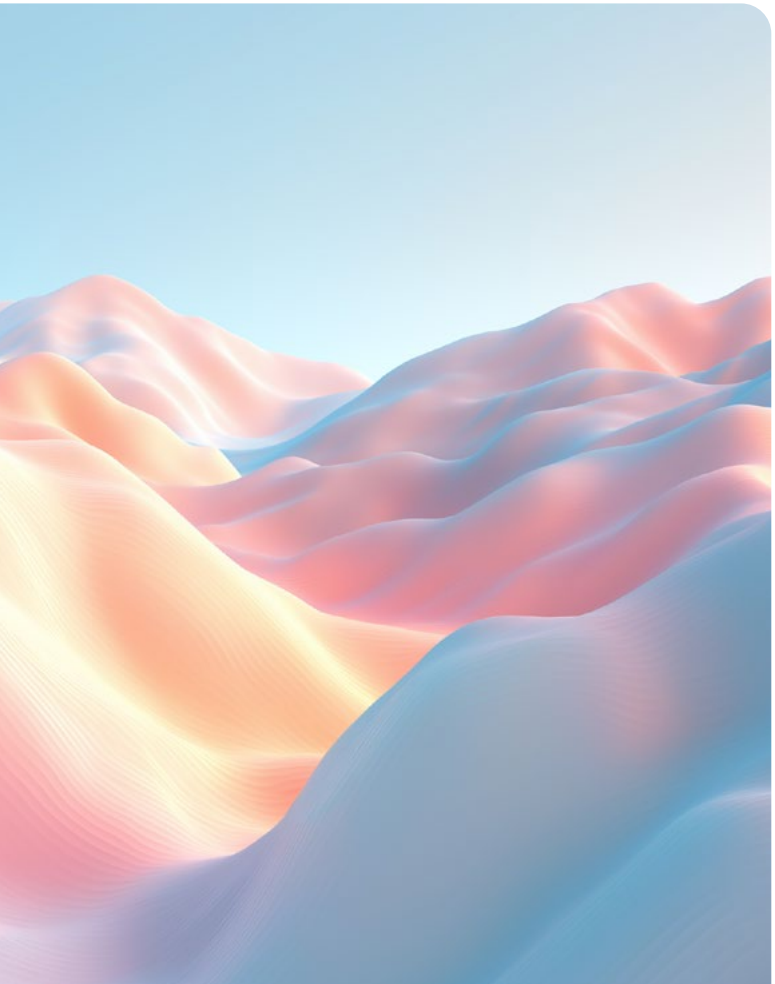
This creates a highly competitive market, driving innovation and providing businesses with a wider range of options to choose from[29].

Beyond Accuracy: The Rise of Efficiency, Compact Models, Improved latency and Advanced Reasoning in AI

While model accuracy on benchmarks remains the most widely scrutinized metric, growing efforts are being devoted to efficiency through model size reduction, context awareness, reasoning capabilities and even agentic capabilities. These factors have a direct impact on operational costs, deployment flexibility, and the ability of these AI models to address complex problems.

Model Distillation: Enabling compressed SOTA AI models

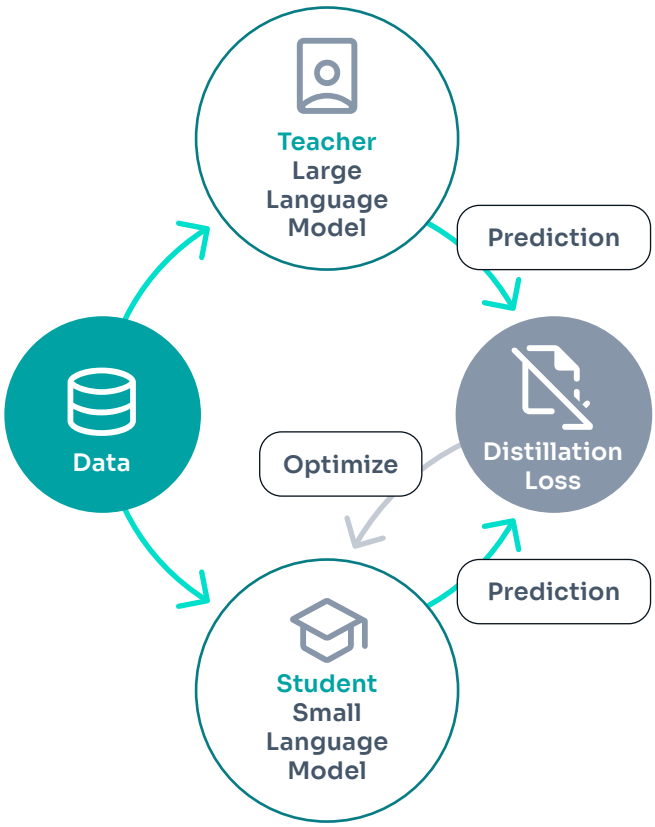
Distillation is a technique that enables the creation of more compact and efficient models by transferring knowledge from a large "teacher" model to a smaller "student" model, preserving performance while reducing size and computational cost.



In early 2025, DeepSeek released DeepSeek R1, a powerful 671B-parameter language model, alongside its distilled versions : R1-Lite (7B) and R1-Mini (1B), while R1 excels in complex reasoning tasks such as code generation and mathematics, R1-Lite achieves comparable performance on summarization and question-answering.

Distillation works by having the student model mimic the teacher's softened output probabilities (logits) or intermediate representations, forcing it to learn nuanced patterns rather than just hard labels. This results in more compact and faster models ideal for edge devices.

FIGURE 12: KNOWLEDGE DISTILLATION PRINCIPLE: TRAINING A SMALLER MODEL USING A LARGER TEACHER MODEL



Model Quantization: Pushing the Boundaries of Efficiency

Quantization is a deep learning technique that reduces the computational and memory demands of LLM by converting high-precision numerical representations (e.g. 32-bit floating point)

(29) Chatbot Arena (formerly LMSYS): Free AI Chat to Compare & Test Best AI Chatbots

into lower-precision formats (e.g. 8-bit integers). This optimization lowers resource requirements for inference, maintaining acceptable accuracy and making scalable LLM deployment possible. By shrinking model sizes without significant performance loss, quantization allows businesses to run powerful AI on cost-effective hardware, enhancing real-time AI applications for both cloud and edge environments.

In fact, quantization can even be used to reduce training costs. For example, the Deepseek-V3 model's training leveraged an FP8 (8-bit Floating Point) mixed-precision training framework to improve cost-effectiveness. More recently, Microsoft's research into 1-bit quantization (e.g., 1.58-bit with ternary parameters) has further advanced this technique, demonstrating its potential to achieve even greater efficiency without compromising performance[30]. This suggests a future where

> “Powerful models will likely operate with greater efficiency enabling their integration into embedded GenAI applications.”

The emergence of new model innovations

In the pursuit of greater efficiency, new model architectures are emerging. Two notable examples are Hybrid Models and Mixture-of-Experts (MoE) models:

HYBRID MODELS

These models combine LLMs with State-Space Models (SSMs), such as Jamba 1.5[31] , to achieve higher throughput and longer context windows. SSMs are dynamic systems that use state variables and equations to model their evolution over time. They are emerging as a potential solution to LLM hallucinations and can achieve state-of-the-art performance while processing larger amounts of data more efficiently

MIXTURE-OF-EXPERTS (MOE)

Models built on this architecture activate only the most relevant "expert" sub-models for specific tasks, saving computational resources. This translates to lower energy consumption and faster processing, reducing operational costs. While MoE models offer faster pretraining and inference than dense models of comparable size, they demand high VRAM, as all experts must be loaded into memory.

On December 8, 2023 Mistral released Mixtral 8x7B, an open source sparse MoE, which matched or outperformed GPT-3.5 on most benchmarks. This was an indicator of what was to come, with the release of the DeepSeek V3 MoE model on December 26th, 2024, which outperformed LLAMA-3.1, GPT-4o, and Claude-3.5-Sonnet on multiple benchmarks[32].

Despite DeepSeek having 671B parameters, only 37 billion of them are activated for any task, reducing computational costs without loss of performance thanks to MoE. Alibaba Qwen’s MoE innovations - such as fine-grained experts, sparse activation, and advanced routing - enable their models to deliver state-of-the-art performance at a fraction of the cost and computational requirements of traditional dense models. These improvements deliver up to 75% reductions in training costs, nearly double the inference speed, and the ability to outperform much larger models on key benchmarks. These are good examples among a variety of innovations in AI,

> “New model architectures are getting more and more mature. It is therefore interesting to see the community leveraging each other’s innovations through open source developments - as Deepseek or Qwen did with Mistral and vice versa probably. ”

(30) The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits
(31) Lieber, O., Lenz, B., Bata, H., Cohen, G., Osin, J., Dalmedigos, I., ... Shoham, Y. (2024). Jamba: A hybrid transformer-mamba language model.
(32) LLAMA-3.1-405B-Instruct, GPT-4o-0513, and Claude-3.5-Sonnet-1022 – Notably on MMLU-Pro, Math 500 and Codeforces.

Test Time Scaling or “Reasoning”: Solving More Complex Problems

The original scaling laws in the context of Large Language Models (LLMs) primarily refer to the pre-training phase. The fundamental principle is that larger models, trained on larger datasets with greater computational resources, generally exhibit greater capabilities. In essence, LLM performance scales with model size, dataset size, and the amount of compute used during training.

The next set of scaling laws pertain to the post training phase, using techniques including fine-tuning, distillation, and reinforcement learning. This type of scaling can improve computational efficiency, accuracy or domain specificity[33].

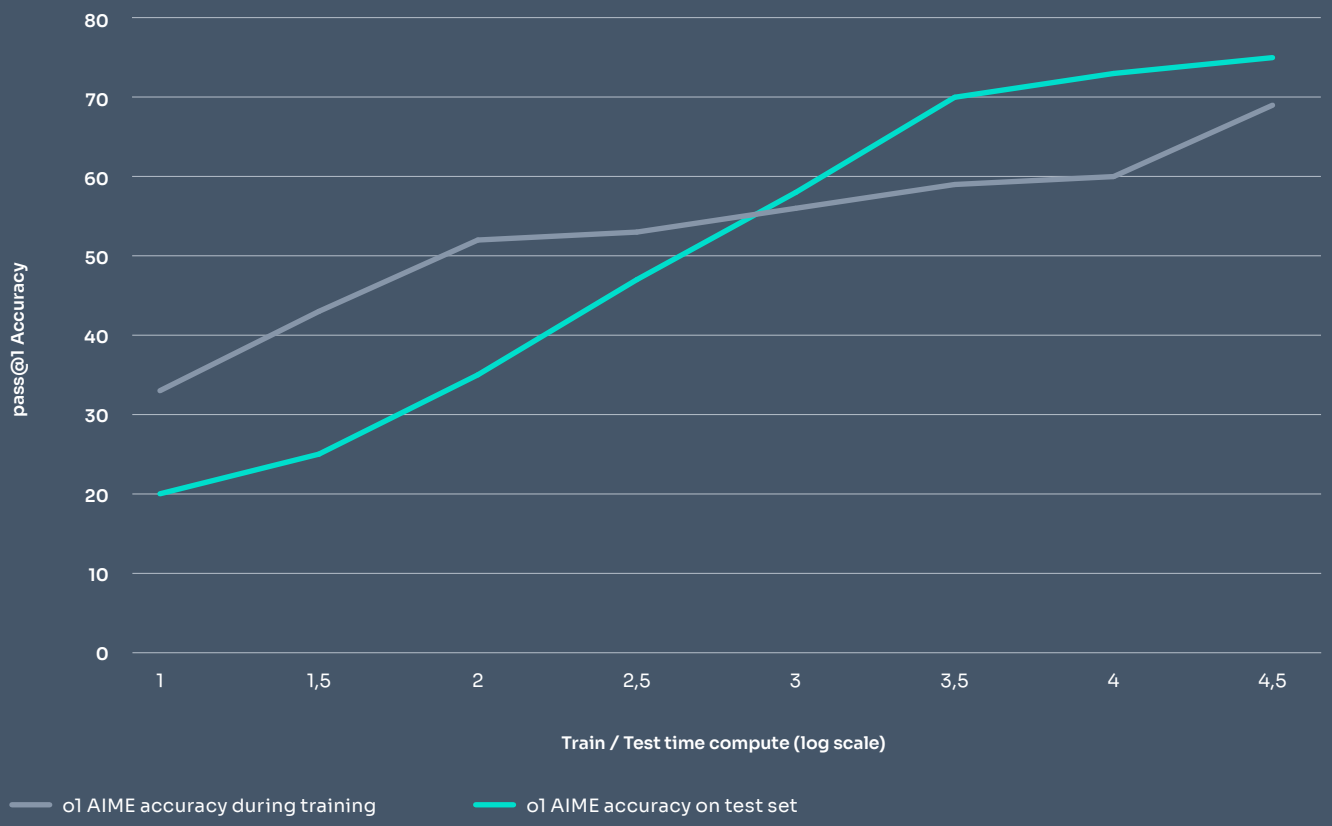
OpenAI pioneered models leveraging the newest scaling law, or the concept of "test-time scaling." This approach, while potentially less intuitive for machines, aligns more closely with human problem-solving. The underlying principle is that

> “Providing an AI model with more processing time to "think", results in more accurate and nuanced outputs.”

The development of "reasoning LLMs," such as GPT-o1, GPT-o3, and DeepSeek R1, exemplify this trend. While Claude 3.7 Sonnet is also a reasoning model, it differs by employing a user-definable "thinking token budget." This allows users to enable or disable "thinking" and adjust the number of tokens allocated to it, up to a maximum of 128,000 output tokens (with a default of 1024). higher risks of hallucinations.

FIGURE 13: THE EFFECT OF TEST-TIME COMPUTE ON ACCURACY

Source: OpenAI o1 Report



(33) https://blogs.nvidia.com/blog/ai-scaling-laws/

Reasoning models are built on top of foundation models and consistently outperform them across nearly all benchmarks. The o-series (o1 & o3) were (most likely) developed on top of GPT-4o, though this has not been confirmed. GPT-4.5 demonstrated a 20-percentage-point improvement in GPQA[34] scores compared to GPT-4o, suggesting that next-generation reasoning models will likely follow this trend, pushing the boundaries of existing benchmarks and requiring the development of new metrics.

For businesses, these enhanced reasoning capabilities enable AI to move beyond basic-level automation to tackle more complex, advanced tasks. With reasoning models and improved tool use, AI can now handle broader use cases that were previously out of reach using traditional LLMs:

Multi-hop Q&A:

Connecting layered facts (e.g., "What were the key recommendations from the Q1 market analysis report, and how do they align with our current product development roadmap?").

Math/Logic:

Solving word problems with step-by-step derivations (e.g., "If Alice is twice as old as Bob was 5 years ago...").

Code Debugging:

Fixing logic errors, not just syntax (e.g., "This function fails due to uninitialized variables").

Legal Nuance:

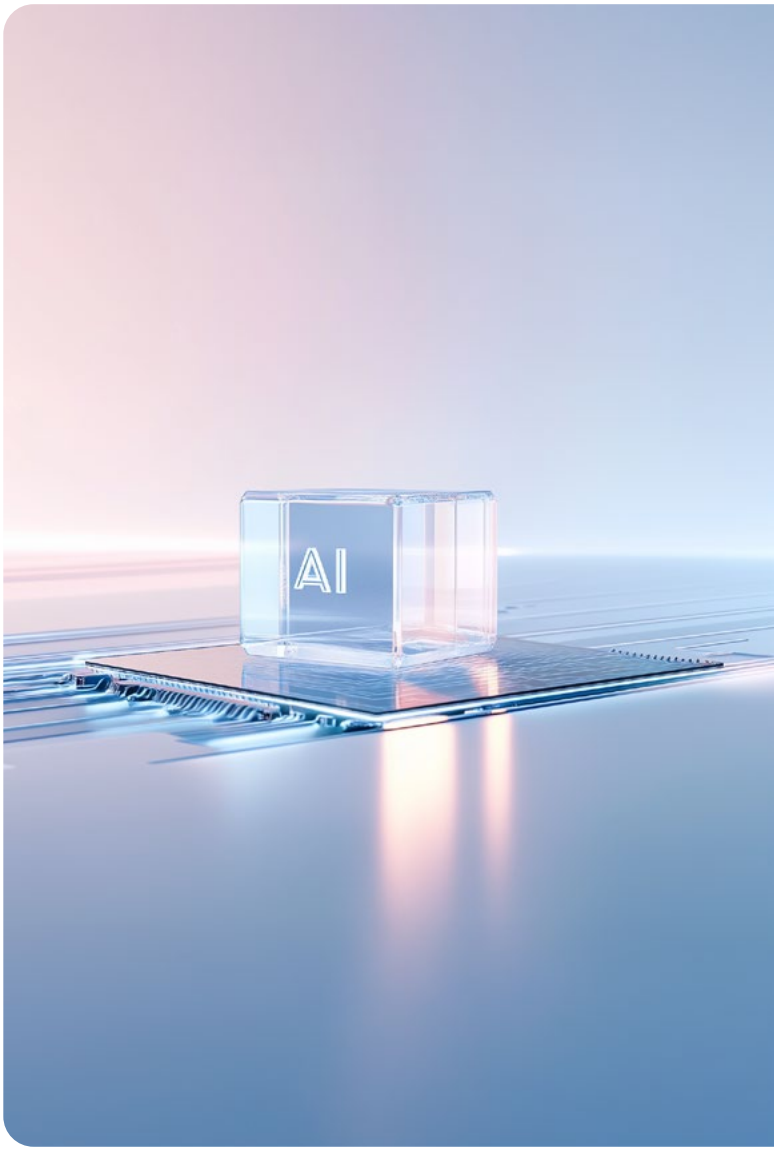
Interpreting clauses contextually (e.g., "Does 'force majeure' cover blockchain outages?").

Cross-Domain Insight:

Merging fields (e.g., "Explain quantum computing using cooking analogies").

Chain-of-Thought prompting, RAGs, and logic-trained fine-tuning enable these leaps and offer deeper insights and better decision-making capabilities. Even if they don't always dominate traditional benchmarks,

> “Reasoning LLMs are fundamentally changing what AI can do for businesses as it improves model autonomy.”



(34) Graduate-Level Google-Proof Q&A Benchmark, comprising domain expert level questions in biology, physics and chemistry. PhDs in their domain score only around 65%

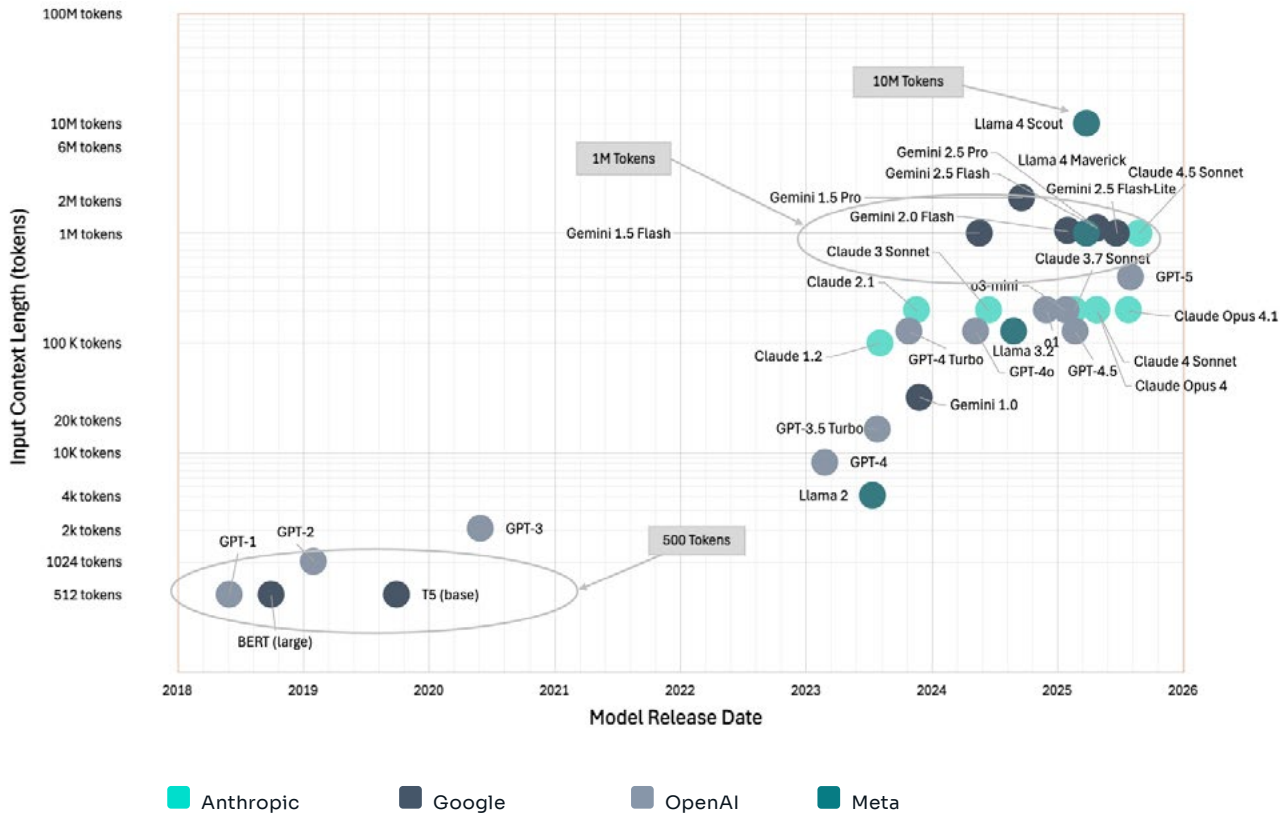
The exponential growth of context length in LLMs

Higher quality data leads to better model performance

The aggregate score is the aggregated evaluation score of LLMs on various benchmarks, evaluated using the lighteval library.
The legend mentions various internet scale datasets will several billion tokens, that are used to train LLMs. To better compare the datasets, we plot a 3-point moving average (MA) for all the datasets considered.

FIGURE 14: EXPONENTIAL GROWTH OF CONTEXT LENGTH

Source: Artfish AI: Evaluating long context large language models



Complementing the advances in model architecture, quantization, and test-time scaling, ongoing research is focused on expanding the context window of LLMs. Think of an LLM's context window as its "working memory." This is the amount of information it can consider at once, analogous to human short-term memory. Google's Gemini model series, with its >=1 million token context window, exemplifies this approach. The goal is to potentially eliminate the need for a separate vector database (used in RAG) by allowing the LLM to directly process much larger amounts of information. However, simply increasing the context window introduces a "needle in a haystack" problem: models often struggle to effectively retrieve specific information from extremely large contexts, necessitating further research into efficient information retrieval mechanisms.

While it may require additional initial investment to setup a RAG, it will likely be cheaper in the long run. Users will probably end up paying more for large context models since the cost is calculated based on both input and output tokens[35]. The advantages of RAG systems are detailed further in **Section IV : Building Effective Systems**.

(35) https://blog.dataiku.com/is-rag-obsolete

Agentic AI: unlocking actions and tool usage

The What, Why and How of AI Agents

Agentic AI represents a significant leap beyond single-model AI systems. Instead of relying on a single LLM, agentic AI involves systems composed of language models that interact with each other or with third-party products and tools.

These interactions are not simple question-and-answer exchanges; they encompass a wide range of tasks, including planning actions, reviewing outputs, simplifying inputs, or setting new goals. This enables the creation of complex chains of thought, where models interact with multiple systems, empowering them to use these third-party products or tools to solve problems that single LLMs struggle to correctly address.

FIGURE 15: DIFFERENCE BETWEEN RPA, AI, GENAI AND AI AGENTS

	Robotics process automation	Traditional AI	GENERATIVE AI	Agentic AI
WHAT IS IT GOOD FOR?	Unique machine learning algorithm. More structured and constrained than generative AI.	Pattern recognition, regression analysis/prediction, classification.	Content generation (e.g.text, images, code, etc.)	Decision-making and autonomous action, Inter-system interactions.
LEARNING	Rule-based imitation. No learning phase.	Unique machine learning algorithm. More structured and constrained than generative AI.	Self-supervised, unsupervised, representation in latent space.	Reinforcement learning, unsupervised learning.
USE CASES	Task automation, data entry, process automation.	Human augmentation, customer segmentation, predictions.	Text, images, audio, code generation.	Personalized assistants, autonomous AI, and collaborative work.
BENEFITS	Operational efficiency. Cost reduction.	Decision support Optimization.	Customization Creativity.	Real-time and collaborative interaction, Autonomy, Improvement of complex processes.

A typical workflow follows a "Thought-Action-Observation" cycle: the agent decomposes a complex problem, takes specific actions using available tools (APIs, etc.), observes the results, and adapts its strategy as needed. Agents can also learn from experience through memory, refining their problem-solving abilities over time.

Several open frameworks exist for building and monitoring AI Agents, including Smolagents (HuggingFace), LlamaIndex (Meta), LangGraph (LangChain). These open frameworks provide integrations with hundreds of tools for web browsing, code interpretation, productivity applications like GitHub and Jira, and database access. GUI-based tools like Rivet and Vellum[36] further simplify the agent building process.

(36) https://www.anthropic.com/engineering/building-effective-agents

By orchestrating these tools, AI Agents can tackle an ever-expanding range of real-world challenges. Proprietary frameworks, integrated in larger cloud services also exists such as Amazon's Bedrock Agents for building applications with company-specific APIs and systems, CrewAI for multi-agent systems, Agent Squad for orchestration; Vertex AI Agent Builder for developing, deploying, and monitoring generative AI agents on GCP ; OpenAI Agents SDK for designing multi-agents workflow with OpenAI models and supported on Azure environments.

Similarly, platforms like n8n, a workflow automation platform, provide a blend of no-code and code-based approaches, allowing technical teams to build powerful automations with a wide range of integrations and native AI capabilities. Another approach is Manus, a general AI agent designed to execute tasks, offering pre-built use cases and a focus on turning user intentions into concrete actions. These tools cater to different skill sets and needs, making AI agent development more accessible to a wider audience.

The Next Generation of Robotic Process Automation

AI Agents can be seen as the AI evolution of traditional Robotic Process Automation (RPA). While RPA relies on predefined scripts for repetitive tasks, AI Agents reason, plan, and adapt dynamically, unlocking new automation possibilities.

Key advantages over traditional RPA include:

Intelligent Data Extraction:

Extracting data from unstructured sources (emails, PDFs) with greater accuracy. RPA fails to adapt to unseen formats.

Dynamic Workflow Automation:

Adapting to changing requirements without constant human intervention.

Self-Improvement:

Learning from experience to improve performance.

Integration with Legacy Systems:

AI Agents can use classical tools such as accessing files in a file folder, automating mouse and keyboard and other non-API tools to be connected to legacy systems.

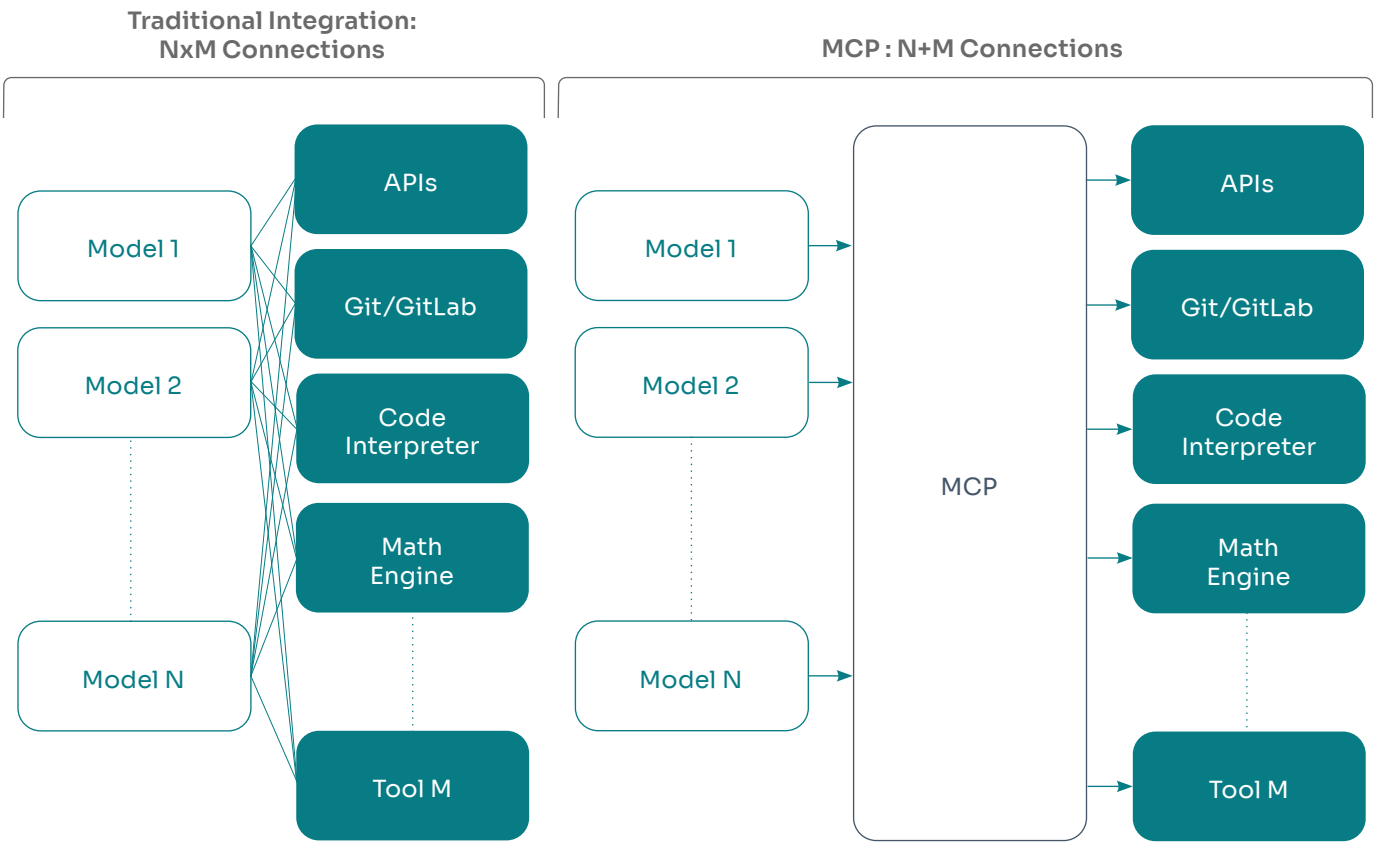
AI Agents can go even further and even optimize the tool integrations that one has to manage. While current agent frameworks provide access to hundreds of tools, connecting these tools to a common framework for seamless use is something that needs to be handled. This is where Model Context Protocol (MCP) comes in.

Model Context Protocol (MCP)

If you work in Tech, there is a good chance you might have heard about MCP. MCP aims to simplify the connection of AI to tools and data sources, first introduced by Anthropic in the last quarter of 2024 as the "USB-C port" for AI. MCP provides a universal and open-source standard to connect AI to tools and data sources. MCP converts an MxN tool integration in LLM problem into an M+N problem. Instead of having to connect to each of the N models to M different tools, (=MxN integrations), we can simply connect the models to the MCP server (N connections) which will in turn be connected to M different tools.

In MCP terms, a client is a software like Claude Desktop or Cursor that a user interacts with directly, and which incorporates an LLM and grants it access to tools provided by MCP servers (see **Figure 17**). MCP servers are thus programs you install and run on your own machines.

FIGURE 16: BENEFITS OF USING MCP OVER TRADITIONAL INTEGRATION METHODS
Source: Hugging Face Blog: "What Is MCP, and Why Is Everyone – Suddenly!– Talking About It?" & Anthropic



While libraries such as LangChain allow developers to equip LLMs with tools, MCP is model-centric, allowing a running agent to discover and utilize any MCP-defined tool at runtime, reducing the need for pre-built integrations. For developers and businesses, this translates into faster AI deployments and lower development costs. By allowing AI assistants to seamlessly access data, automate complex tasks with tools, and generate answers without complex custom integrations, MCPs unlock scalable AI solutions in any IT organization.

MCP's innovation can be compared to APIs (Application Programming Interface) in terms of impact on digital systems and with improved capabilities in terms of integration, scalability and fault isolation. MCP is to AI what API is to data, both technologies enable integration and communication between digital systems. Most digital environments of the future will likely use both: APIs for structured connectivity, and MCP as the intelligent, adaptive layer for AI-driven workflows. See MCP vs API illustration below.

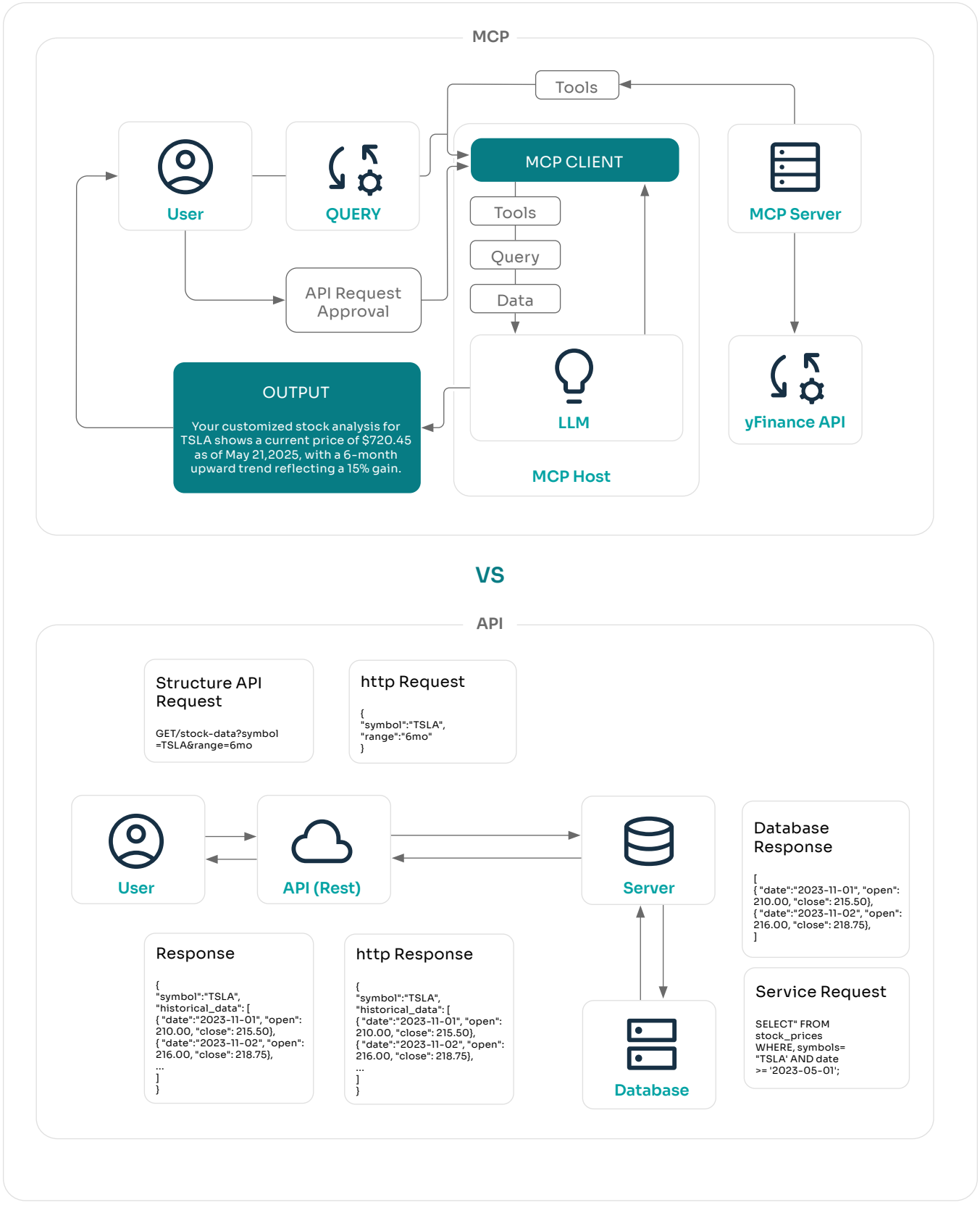
Although this protocol still needs to prove its maturity, especially in terms of security (e.g. prompt injections), it is worth noting that all the big actors (including OpenAI and Google) are already adopting it as a common protocol.

Google recently released another framework called **A2A - Agent to Agent** to simplify AI agents' collaboration, this framework also uses MCP. Though promising, this framework is niche and still needs to prove itself valuable in production environments. The reasons remain the same : AI agents can and will make mistakes. It is however worth mentioning as these technologies are only a few months old, new improvements should progressively resolve these limitations.

Table 1: The key differences and similarities between API, MCP and A2A protocols

Aspect	API	MCP	AGENT TO AGENT
PURPOSE AND SCOPE	General system-to-system communication; exposes fixed endpoints	AI-to-tool integration; standardizes how AI models access tools & data	AI agent-to-agent communication; enables collaboration between autonomous agents across different platforms and frameworks
INTERACTION MODEL	Stateless, request-response; each call is independent	Stateful, bidirectional; supports real-time, continuous context-sharing	Task-oriented, stateful communication; supports both stateless request-response and long-running collaborative tasks with real-time updates
INTEGRATION APPROACH	Custom integration for each system; many bespoke connections	Universal, discovery-based; single protocol for dynamic AI interactions	Agent discovery via "Agent Cards"; open protocol enabling agents from different vendors/frameworks to interoperate seamlessly
DATA HANDLING	Returns structured data (JSON, XML); context managed externally	Context-aware; aggregates/formats data for AI models dynamically	Supports multimodal data exchange (text, audio, video); structured messaging with parts and artifacts
SCALABILITY & FLEXIBILITY	Scaling often means scaling the whole system; less flexible	Microservices-based; scale/update individual services independently	Microservices-oriented; agents act as independent services that can be swapped, updated, or added with minimal friction
FAULT ISOLATION	Failures can affect entire system	Faults isolated to individual services	Failures isolated to individual agents; protocol supports task retries and backup agent routing for resilience

FIGURE 17: MCP VS API



/ Chapter 3 – Table of contents

Navigating the AI Landscape

3. Navigating the AI landscape

The Generative AI landscape is interconnected, with technological advancements intertwined with geopolitical ambitions, market forces, and regulatory developments. This section explores this complex web, examining how the global race for AI dominance influences market leadership and hardware supply chains, drives significant investment, and shapes the trajectory of major players like China. Furthermore, we will highlight the critical importance for businesses to understand the nuances of data partnerships and the diverging legal and regulatory frameworks that are defining the boundaries of AI deployment and requiring careful strategic consideration

Growing geostrategic stakes

Having examined the fundamental shifts in the AI landscape, we now see AI and especially GenAI emerging as a critical geostrategic asset with a growing influence on international relations and policy decisions.

The United States, at the forefront of AI innovation, provides compelling examples of this evolving dynamic through several recent strategic actions:

Securing AI dominance:

- The United States required TSMC to establish a semiconductor manufacturing facility in Arizona as a strategic decision, considering that semiconductor production is concentrated among a small number of Taiwan-based companies. Given China's interests regarding Taiwan, developing domestic chip manufacturing capability is essential for the U.S. to maintain technological self-sufficiency.
- Europe, through the European Chips Act, aims to secure semiconductor supply by mobilizing over €43 billion in public and private investments. This initiative strengthens Europe's technological leadership and enhances supply chain resilience through collaboration with Member States and international partners. It also aims to increase Europe's global semiconductor market share to 20% by 2030.

Controlling foreign interactions:

- The U.S. blocked a Saudi fund's partial acquisition of Anthropic
- The U.S. restricted Chinese access to Nvidia hardware, limiting China's AI infrastructure development, though companies like DeepSeek demonstrate that Chinese firms can still access U.S. chips obtained before the limitations or rent data centers to work around these limitations.

Deepening AI's role in national security:

- Retired U.S. army general Paul M. Nakasone joined OpenAI's board.
- Anthropic and Palantir recently partnered, allowing Palantir to utilize Claude models[37].
- Google removed its commitment not to develop AI for weapons and surveillance from its corporate policy, reversing its previous stance against technologies causing harm[38] .

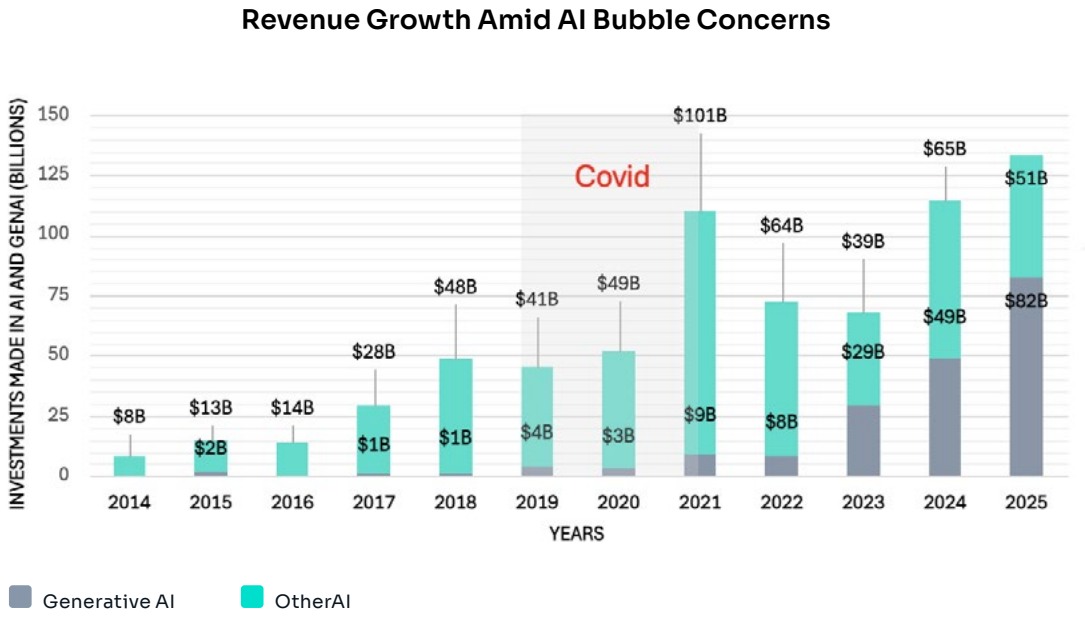
AI will play a major role in the future balance of power among global leaders.

(37) Writer, R. L. C. (2023, December 8). Employees are feeding sensitive Biz data to ChatGPT, raising security fears.
(38) Lesnes, C. (2024b, September 30). California Governor Gavin Newsom vetoes AI safety bill. Retrieved from https://www.lemonde.fr

Market Leadership and Hardware Dominance

/FIGURE 18:GENAI INVESTMENTS HAVE STRONGLY INCREASED OVER THE PAST TWO YEARS
Source: State of AI Report 2024

GenAI investments have significantly increased over the past two years
Thank to GenAI applications and products, and mega-rounds of investment into companies like OpenAI, xAI and Anthropic, investment into AI companies reached close to $100B. It might have even exceed it since the time of writing.



As generative AI becomes a key geostrategic asset, it is also reshaping market dynamics, with a few dominant players capturing most of the value. This revenue concentration mirrors the geopolitical race for AI leadership and reveals growing tensions between innovation, valuation, and financial sustainability.

The growth of the GenAI industry has led to significant valuations across leading companies. OpenAI's $157 billion valuation in October 2024 stands in stark contrast to its $3.7 billion revenue and $5 billion annual losses[39]. As the market leader, OpenAI generates nearly four times the revenue of Anthropic, its closest competitor.

In Europe, four of the five largest tech fundraising rounds in 2024 were AI-related: Wayve ($1.05 billion), Mistral (€468 million), Poolside ($500 million), and Helsing (€450 million)[40] . This strong momentum shows that AI is also a strategic priority on the continent, with growing ambition to shape the global landscape.

While increased investment drives innovation and competitive pricing now, it may lead to market instability later. Companies face mounting pressure to develop sustainable business models in this rapidly changing landscape. The gap between current valuations and financial realities raises legitimate concerns about a potential AI bubble.

Observers legitimately question whether these valuations reflect realistic growth potential or speculative enthusiasm. The market may eventually require significant consolidation as investors demand clearer paths to profitability.

Nvidia's continued dominance in AI hardware

Beyond its innovation breakthrough, the rapid rise of GenAI in the market is largely due to significant advances in hardware, with leaders such as Nvidia playing a central role in the emergence of these technologies.

(39) OpenAI sees roughly $5 billion loss this year on $3.7 billion in revenue
(40) The 20 largest equity and debt rounds in 2024

Nvidia's leadership in the AI hardware market is undeniable, reflected in its exceptional financial performance. In Q2 2024, the company generated $30 billion in revenue, substantially exceeding AMD's $5.8 billion and Intel's $12.8 billion as seen in **Figure 19**. This advantage highlights Nvidia's strategic position as the leading provider of AI accelerators.

In addition to their top-tier hardware, such as the flagship H100 chips, the company's comprehensive CUDA ecosystem strengthens its market position by offering seamless integration with major machine learning frameworks, creating significant barriers to entry for competitors. This strategic integration of hardware and software promotes user dependency, as developers and researchers are encouraged to adopt the entire Nvidia ecosystem, from GPUs to the CUDA framework. This approach effectively locks users into an environment where the transition to other platforms becomes both technically and financially onerous.

In contrast to competitors focusing on specialized chips, Nvidia's early lead allowed it to develop versatile GPUs capable of handling diverse AI workloads. This strategy enables the company to address requirements across multiple domains, from generative AI to reinforcement learning, effectively capturing demand across the AI landscape.

While this dominance demonstrates Nvidia's successful market strategy, it also raises concerns about technological dependency and potential limitations on innovation in AI hardware development.

FIGURE 19 : 2024 Q2 REVENUE AND EARNINGS PER SHARE OF NVIDIA, AMD AND INTEL
Source: State of AI Report 2024

Q2 2024 Revenue & Earnings/Share for major AI hardware companies

Alternatives beyond NVIDIA

Competition in the AI hardware market is intensifying, while NVIDIA remains dominant, several major players are developing compelling alternatives. AWS introduced the Inferentia and Trainium chips for cost-effective AI inference and training, with Trainium2 offering 30–40% better price-performance than general-purpose GPUs, making it a compelling alternative for large-scale training workloads.

Google's TPUs and ARM-based processors such as Ampere Computing also provide powerful alternatives, diversifying the AI hardware landscape. These initiatives reflect a broader strategic shift: companies are increasingly designing and adopting specialized architectures that move beyond the limitations of general-purpose GPUs.

This enables higher performance per watt and improved cost control, while also aligning infrastructure with specific AI use cases. Alternatively, inference performance, often constrained by memory bandwidth rather than raw compute power, has become a major focus for innovation.

Cerebras's WSE-3[41] integrates 44GB of SRAM directly on-chip, enabling ultra-fast token generation, with speeds of up to 1,800 tokens per second for Llama 3.1 8B, compared to 242 tokens per second on an H100. Similarly, Groq's LPU takes also an SRAM-based approach, delivering high-speed inference with reported performance around 250 tokens per second.

Google's TPUs and ARM-based processors such as Ampere Computing also provide powerful alternatives, diversifying the AI hardware landscape. These initiatives reflect a broader strategic shift: companies are increasingly designing and adopting specialized architectures that move beyond the limitations of general-purpose GPUs.

This enables higher performance per watt and improved cost control, while also aligning infrastructure with specific AI use cases.

As AI hardware diversifies, businesses stand to benefit from greater flexibility in AI deployment, cost optimization, and reduced vendor lock-in. With alternatives tailored to specific workloads, companies can select hardware best suited to their needs rather than relying solely on NVIDIA's ecosystem. This growing competition is also expected to drive price-performance improvements, making AI more accessible for a wider range of enterprises.

Skyrocketing investments in computing infrastructure

The demand for advanced AI capabilities is driving investments in computing infrastructure. The Stargate Project, backed by SoftBank, OpenAI, Oracle, and MGX, outlines $500 billion in private investment over the next four years, with an initial $100 billion deployment[42].

Meanwhile, the European Union's InvestAI initiative is set to mobilize €200 billion, including a €20 billion fund for AI "gigafactories" - public-private research hubs designed to train ultra-large AI models on clusters of 100,000+ next-gen chips[43].

For businesses, these initiatives signal explosive growth in AI infrastructure, creating opportunities in data center construction, hardware, and AI services. The emphasis on public-private partnerships offers new avenues for collaboration, while the geographic concentration of investments underscores growing geopolitical competition in AI. Additionally, the scale of these projects translates into an increasing demand for suppliers across the AI ecosystem.

However, such massive and energy-intensive data centers have had strong impacts on energy consumption and carbon emissions; in 2022, data centers represented approximately 2% of global electricity usage meanwhile the figure is projected to double by 2026 mainly due to AI[44]. In fact, to power GenAI in 2027 and fulfill the escalating demand for online services and generative AI products, it will be necessary to consume:

- 85 to 134 TWh of electricity representing €1.6 B to €2.6B per year. To put this in perspective, this electricity consumption alone could power a country the size of Ireland (32 TWh[45]) or Portugal (50 TWh[46]) annually or be equivalent to the annual electricity consumption of a major European city like Paris (89.7 TWh[47]) or London (134 TWh[48]).
- 6.6 billion m^3 of water per year for cooling data centers. To grasp the enormity of this volume, it is roughly equivalent to the half the UK's water withdrawal over the course of a year[49].

One striking example of that impact of AI is the case of Microsoft, which pledged in 2020 to be carbon-negative by 2030, and has seen its emission rise by 30% since then, mainly driven by AI data center construction[50]. This trend is set to continue, with Microsoft planning to invest $80 billion in data center infrastructures in FY2025.

This case highlights the tension between technological expansion and environmental goals in the AI era.

(41) Cerebras gives waferscale chips inferencing twist, claims 1,800 token per sec generation rates
(42) Announcing The Stargate Project | OpenAI
(43) Commission launches new InvestAI initiative to mobilise €200 billion of investment in AI | Digital Skills & Jobs Platform
(44) Global data center electricity use to double by 2026 - IEA report - DCD
(45) Ireland annual total energy consumption
(46) Portugal annual total energy consumption
(47) Paris annual total energy consumption
(48) London annual total energy consumption
(49) How datacenters use water - and why kicking the habit is nearly impossible
(50) Microsoft pivots as AI makes emissions cuts toughter

China's AI Industry: Progress Despite Sanctions

Overcoming Hardware Constraints

China's AI industry faces significant challenges due to hardware sanctions and limited access to advanced foreign technologies. For instance, U.S. sanctions prevent Chinese companies from accessing NVIDIA's A100 and H100 GPUs, which are critical for training large language models (LLMs) and running complex AI workloads. These GPUs are considered essential for the most advanced AI systems, and their unavailability poses a major obstacle to China's ambitions in the AI race. The inclusion of the H200 GPU in the blacklist underscores the expanding scope of export restrictions on cutting-edge AI hardware.

Despite sanctions and limited access to foreign advanced hardware, China's AI industry continues to develop rapidly, making notable progress with LLMs. This resilience is due to strategic investments by Chinese technology giants in the development of competitive domestic AI chips.

Huawei's Ascend 910B and Baidu's Kunlun Gen 2 chips, manufactured using 7nm technology, demonstrate China's growing AI hardware capabilities. These chips offer performance comparable to NVIDIA's A100 GPUs, demonstrating China's ability to produce advanced hardware independently of Western sources.

Currently, the hardware capability gap is around three years between NVIDIA's A100 and Huawei's Ascend 910B, and around two years between NVIDIA's H100 and Baidu's Kunlun Gen 2. Importantly, this gap is steadily narrowing, reflecting China's growing ability to bridge the performance gap, marking a decisive shift in the global AI hardware landscape.

A key example of China's innovation within hardware constraints is DeepSeek, which has advanced multi-GPU communication and load balancing for Mixture of Experts (MoE) architectures, improving both intra- and inter-GPU performance. Additionally, DeepSeek has pioneered "pure" reinforcement learning techniques, independently enhancing AI model efficiency. These distinct engineering advancements demonstrate China's strategic move toward technological autonomy, enabling its companies to compete globally despite restricted access to Western technologies.

Growth of Research & AI Companies

China's AI ecosystem is expanding rapidly, fueled by both established tech giants and emerging startups. Key players like Alibaba, Baidu, Tencent, and ByteDance, alongside newer startups such as DeepSeek, Zhipu AI, and Baichuan, are making significant contributions to AI research and application development.

The following table summarizes key AI companies, their valuations, and funding, highlighting the diversity and strength of China's AI ecosystem.

Company	Valuation	Funding & Comments
Baidu	$31.46 billion	>$101 million, IPO completed, est. 2000, well-known and has
Tencent	$589.03 billion	been around for a long time. Not a startup anymore.
SenseTime	$54.40 billion	Not a startup anymore.
4Paradigm	>$2 billion	$2.82 billion, 9 rounds, IPO completed, est. 2014
Yitu Technology	$3.5 billion	>$230 million, IPO completed, est. 2014
Baichuan	>$1 billion	>$385 million, series C
Zhipu AI	$2.74 billion	>$700 million
Moonshot AI	$2.5 billion	>$206 million, $137M founding, $69M Series D
MiniMax	$2.5 billion	>$1 billion
iFlytech	$111 billion	>$600 million
CloudWalk Tech.	$2.18 billion	>$407 million, est. 1999
DJI	$15 billion	>$379 million, series B

Despite hardware constraints, Chinese AI companies are implementing innovative technical solutions to advance their AI capabilities. As an example, 01.AI is enhancing Chinese language datasets to close the quality gap with English-language training corpora.

Chinese AI models have now reached competitive parity with leading Western counterparts across multiple domains. Alibaba's Qwen family of models demonstrates performance comparable to GPT-4 and Claude in general capabilities, while achieving notable domain-specific superiority. Specifically, Qwen2.5-Coder has surpassed GPT-4o on the LiveCodeBench evaluation, and Qwen2-VL-72B has attained the top position on the open-VLM[51] leaderboard for vision-language tasks. In the reasoning domain, DeepSeek R1 has emerged as a direct competitor to OpenAI's o-family (o1, o3... o4) in mathematical problem-solving and logical reasoning capabilities.

> In particular, DeepSeek amazed the AI community with a combination of timing, performance, and transparency”

DeepSeek's technical innovations, including its MoE architecture implementation, computationally efficient BF16[52] training protocols, and aggressive pricing strategy at $0.14 per million tokens for DeepSeek V2, have substantially disrupted market dynamics, compelling established providers such as Alibaba Cloud to recalibrate their pricing structures to remain competitive.

Beyond established technology corporations, a new generation of "AI tigers" is emerging in the Chinese ecosystem. Companies including Baichuan, Zhipu, Moonshot, and MiniMax, alongside innovative startups like 01.AI, Stepfun, and Infinigence AI, are driving advancements in fundamental models, synthetic data generation, artificial general intelligence research, and hardware-software integration.

> Collectively, these developments illustrate China's accelerating momentum and strategic ambitions in the global artificial intelligence landscape.”

Data Partnerships, Regulatory and Legal Challenges

Data Partnerships: Navigating Legal Ambiguity in AI Training

AI companies are increasingly relying on their datasets, while at the same time facing increasing scrutiny over their use of proprietary content. The main problem lies in the lack of transparency: creators have minimal visibility into how their work feeds AI systems after being modified, integrated into massive training corpora and obscured in the results.

This opacity has triggered numerous legal complaints from artists, publishers, and educators concerned that their work is being used without consent, attribution, or compensation. The ambiguity around what constitutes "transformative use" only heightens tensions as stakeholders debate whether AI outputs are truly transformative or merely derivative.

To mitigate these risks, major tech companies are forming data partnerships. OpenAI, for example, has signed a $250 million deal[53] with News Corp for access to archives such as The Wall Street Journal and Barron’s. It has also partnered with Le Monde for selected data with attribution and with The Financial Times for compensated access to their content. However, such arrangements remain an exception rather than a rule, leaving many independent creators and smaller organizations excluded from these negotiations.

The legal landscape is rapidly evolving in response. The lawsuit brought by the New York Times against OpenAI and Microsoft in 2023 represents a turning point[54], as it addresses fundamental questions about consent, ownership and monetization in the development of AI.

(51) Huggingface’s open VLM leaderboard (Q1 2025)
(52) BF for Brain Floating Point format, 16-bit floating points, this methodology maintains numerical stability while halving memory requirements compared to FP32
(53) A landmark multi-year global partnership with News Corp
(54) The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

Similarly, the U.S. Supreme Court's Warhol v. Goldsmith decision, which establishes that stylistic transformation alone does not automatically constitute fair use, could significantly reshape AI copyright frameworks in the years to come.

Despite OpenAI's temporary win in a separate copyright case (against publishers including Condé Nast) due to a lack of demonstrated concrete harm[55], regulatory scrutiny is escalating.

Emerging Regulations

The regulatory environment for generative AI differs markedly across major regions, reflecting varied governance priorities and strategic interests. The USA has a more fragmented approach, combining the security testing requirements of Executive Order 14110 with state-level initiatives such as California's SB 1047. The bill's eventual veto by Governor Newsom[56] underscores the challenges in drafting and enforcing effective AI regulations. In the absence of comprehensive federal legislation, antitrust action against major players (such as the Microsoft-OpenAI partnership) has become the main regulatory mechanism.

Europe leads with the comprehensive EU AI Act, establishing a risk-based framework that mandates documentation, assessment protocols, and human oversight for high-risk applications, with full implementation expected by 2026. The deliberations on these matters reached a critical juncture at the AI Action Summit convened in early 2025. Different in every way from its predecessors, complementing investment announcements in response to the U.S. Stargate project, discussions went beyond the usual security issues to focus on common governance to encourage the adoption of AI. It highlighted the urgency of addressing ethical and regulatory concerns while not stifling innovation, and above all,

> “
> This summit revealed deep divergences between Europe and the U.S.”

Elsewhere, China uses AI regulation to strengthen state control and advance national goals. The Interim Measures for the Management of Generative AI Services, effective August 15, 2023, require AI systems to align with core socialist values and mandate algorithm registration with authorities. The government is also heavily investing in domestic chip production and algorithm monitoring to reduce reliance on foreign technologies.

Without international coordination, the growing divergence of AI regulations risks creating a patchwork of fragmented standards that could weaken global accountability mechanisms. As more and more countries develop their own rules, the key challenge will be to design regulatory frameworks that not only foster innovation but also protect the public interest in a world increasingly shaped by artificial intelligence.

(55) Knibbs, K. (2024, November 13). OpenAI scored a legal win over progressive publishers - but the fight’s not finished | WIRED Middle East.
(56) Lesnes, C. (2024b, September 30). California Governor Gavin Newsom vetoes AI safety bill. Retrieved from https://www.lemonde.fr

/ Chapter 4 – Table of contents

Building Effective Systems

4. Building Effective Systems

Recent advancements in Generative AI, particularly in areas like Retrieval-Augmented Generation (RAG) and agentic workflows, are providing businesses with powerful tools to build more effective systems and tackle complex problems previously out of reach. However, harnessing this potential requires a responsible approach that understands and mitigates inherent technological limitations and the associated threats for their business. This section provides a practical look at how to approach building these systems, including assessing readiness for agentic AI and defining the necessary components and orchestration strategies, while also providing essential insights into navigating crucial challenges such as model hallucination, cybersecurity risks, and the broader concerns of AI misuse.

RAG: The essential GenAI use case for businesses

As seen previously, RAGs is the go-to use case for GenAI in organizations, its allows LLM to securely access and utilize company-specific information when generating responses (i.e. without a risk of data being stored or used for subsequent model training). This makes AI tools actionable and relevant in the company's context and environment.

RAGs are the most successful GenAI use case for LLMs in companies, and in most cases, it is the fundamental brick for many other GenAI-powered applications. RAGs already deliver substantial value to organizations, but

> The next phase of GenAI evolution, incorporating these outputs directly into business workflows, promises even greater impact. "

The corporate context-specific nature of RAG responses enables a wide variety of approaches, creating highly adaptable workflows. Table below illustrates that diversity.

CONTEXT PROVIDED TO RAG	GENAI USE CASES
Database Structure with: 1. Entities Definition 2. Relationships between tables and entities	Generating SQL queries to answer questions based on that database
CRM Data with: 1. Historical leads data 2. CRM Structure	Ranking in new leads, automating actions
Knowledge Graph with Relations between entities	Detecting unusual relations (like transactions) which could be seen as fraud
Unstructured document Data Base with manually amended documents, industrial & technical drawings, with product references	Chained relations mapping and dependencies: chained products, industrial pieces, chemicals, sourcing contracts, etc.
Corporate code base	AI Coding Assistant with access to the relevant projects code base. Particularly relevant for front-end development.

Although RAG solutions have demonstrated some ROI, RAG will continue to deliver more value when combined with other GenAI approaches. Thanks to its ability to connect corporate databases with LLM's compressed knowledge and intellectual abilities,

> RAG is set to remain a key tool for bringing a company's unique context into any AI workflow."

Agentic workflows: Enhancing Complex Processes Automation

With the foundational role of RAG systems in workflows and the growing agentic capabilities of GenAI models, AI Agents are emerging as a highly strategic avenue for companies, **with RAG systems often serving as a key component**.

Gartner[57] predicts that by 2028, AI Agents will automatically support 15% of business decisions.

However, when talking about agents, their functions and capabilities vary greatly with the maturity of the system and the company implementing it. Indeed, AI agentic systems can go from a simple chatbot with tools to a fully autonomous system able to continuously learn from its experiences (see **Figure 20**).

FIGURE 20: AGENTIC AI SYSTEM MATURITY LEVELS
Source: Sia's Agentic AI Offering

However, while GenAI is still a rapidly evolving field and future innovations may shift current paradigms, there are already best practices and guidelines that should be followed to build effective systems.

Is the company Agentic AI Ready?

Before starting any agentic developments, a company should assess its readiness for AI Agent systems. An Agentic AI Readiness Assessment is based on 6 pillars:

Strategy and Vision:

Evaluate the alignment of AI agents with the company's strategic objectives and roadmap.

Skills and Adoption:

Measure the maturity level of talents, team training, and user adoption of new AI practices.

Business Processes:

Analyze the integration of AI agents into existing processes and their impact on operational efficiency.

Data:

Analyze the quality, accessibility, and governance of data essential for the learning and effectiveness of AI agents.

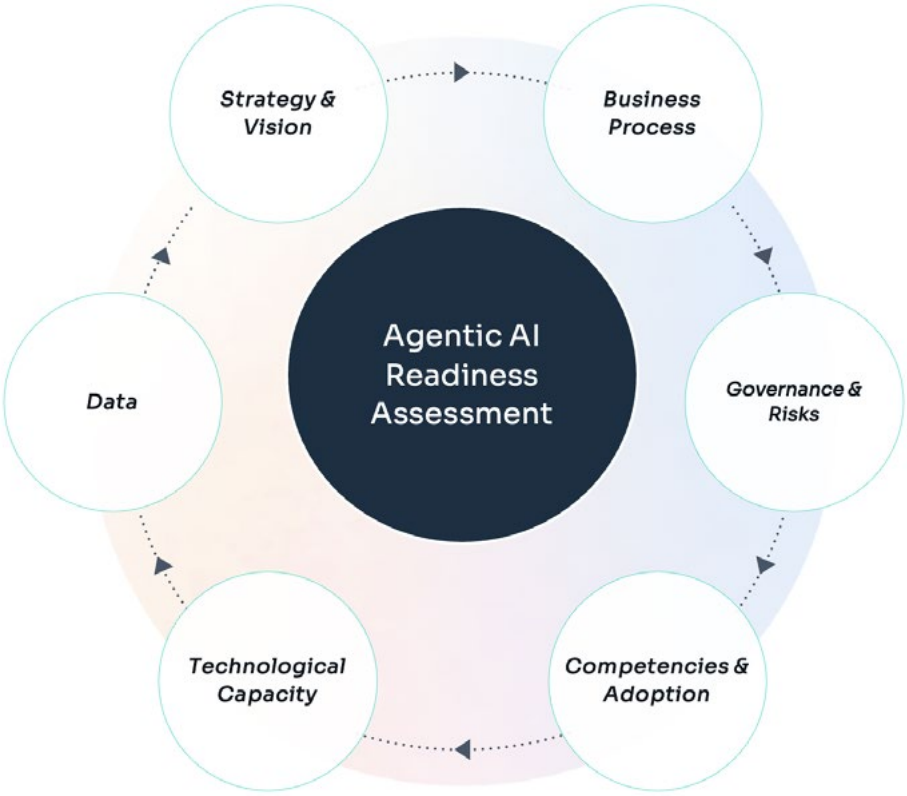
Governance and Risk Management:

Check for the existence of governance frameworks with defined roles and responsibilities, control mechanisms, ethics, and compliance for the responsible deployment of AI agents.

Technological Capabilities:

Assess the infrastructure, tools, and architecture necessary to support the development and execution of AI agents.

/FIGURE 21 : SIA'S AGENTIC AI READINESS ASSESSMENT



When should a company build an AI Agent?

Having assessed the capacity to develop and deploy AI Agent systems, the next step for a company is to identify relevant AI agent use cases.

There is no universal rule for identifying them, but four key aspects can help guide the wayleading to them[58]:

> The next phase of GenAI evolution, incorporating these outputs directly into business workflows, promises even greater impact."

The corporate context-specific nature of RAG responses enables a wide variety of approaches, creating highly adaptable workflows. Table below illustrates that diversity.

Use cases which have resisted automation in the past

Use cases with complex decision-making like decisions based on context

Use cases with rules that are difficult to maintain

Use cases deeply based on unstructured data

Defining the models and tools for your use case

When a use case has been identified, the development of the AI agents can begin.

The foundational form of an agent is composed of:

- **A Language Model** which is the base of all reasoning & decision making.
- **Tools** which are mainly APIs or external functions, extendingfunctional or systems (like RAG) defining the scope of actions of the agent
- **Instructions** which are the guidelines defining how the agent shall behave.

Since these components are fundamental to defining an agent, it is strongly recommended to define them carefully:

- **Choose the most relevant LLM** based on several aspects such as the minimum performance required, the inference speed needed and the acceptable cost for your use case.
- **Define the needed tools** that will:

1. **Collect data efficiently**, using methods including but not limited to Retrieval-Augmented Generation (RAG), web search, etc.
2. **Execute the required actions**, such as sending emails, updating CRM systems, or triggering workflows.
3. **Orchestrate the agent's behavior**, ensuring smooth coordination between data collection, action execution, and decision-making processes.

- **Implement effective instructions** by defining clear actions, breaking down tasks, capturing edge cases, etc.

Determining the best agent orchestration strategy

Having agents is not enough for creating an AI Agent system. It is necessary to orchestrate those agents to create a workflow handling a specific use case.

There are mainly two possible orchestration systems:

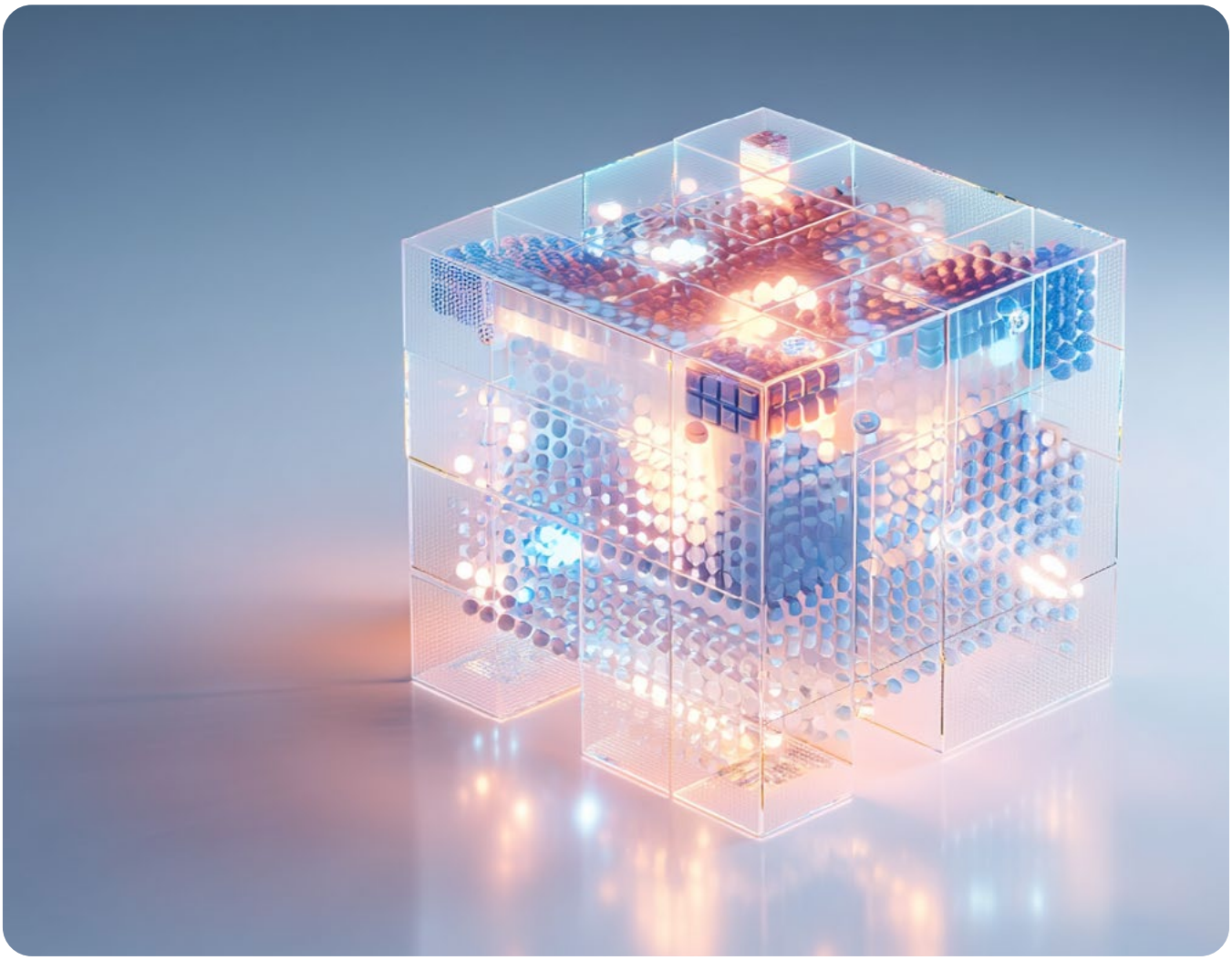
- **Single-agent systems**, where a single model equipped with appropriate tools and instructions executes workflows sequentially.
- **Multi-agent systems**, where workflow execution is distributed across multiple coordinated agents, each with its own function (manager, checker, tool-user, ...).

(58) A practical guide to building agents by OpenAI: https://cdn.openai.com/business-guides-and-resources/a-practical-guide-to-building-agents.pdf

Single-agent systems are well-fitted for use cases with a reasonable amount of logic complexity and tools involved. The multi-agent systems are more appropriate when complexity increases but on the other hand imply a greater challenge in conception and a higher risk of instability.

Setting up relevant Guardrails for your AI use case

To ensure AI agents provide reliable and appropriate responses, it is essential to implement guardrails. These can start with basic rule-based protections or output validation checks. However, as AI agent systems grow more autonomous, guardrail mechanisms are becoming increasingly sophisticated. For example, PII filters prevent the exposure of sensitive personal data, tool safeguards restrict actions like unauthorized financial transactions or mass emails, and proprietary solutions such as Giskard help detect biases or hallucinations before deployment. Additionally, when AI agents utilize external tools or libraries, such as Python packages, the associated risks, like executing malicious code or introducing vulnerabilities, are inherited by the agent. To mitigate these risks, it's crucial to establish guardrails that enforce the use of specific, vetted versions of packages, conduct regular security audits, and monitor for unusual activity. Effective guardrails are typically applied at multiple levels, including input validation, model output moderation, and action execution control.



Sia's top Agentic AI use cases

AI agents will power most future AI applications; here are four of the most valued AI Agentic use cases[59] that we have developed for our business partners:

- **Technology Watch**: An AI agent that monitors and analyzes emerging technologies and industry trends. Intelligent agents gather and filter data from sources like academic publications and industry reports, highlighting relevant insights. This enables organizations to stay ahead of technological advancements, identify innovation opportunities, and make informed strategic decisions.
- **Proposal Generator**: A multi-agent solution where each agent is assigned a specific task. One agent extracts information from RFPs, another selects relevant content from SharePoint, and a third identifies the most appropriate credentials and profiles based on skill matching from the use case catalog. A supervisor agent then consolidates all the collected data into the appropriate repository. This dramatically accelerates proposal creation, freeing up valuable time for consultants to focus on strategic content and client engagement. The result is faster response time on RFPs, improved quality of proposals, and increased win rates.
- **Auto-scraper**: A multi-agent solution designed to assist developers in navigating and scraping web pages more efficiently. Different agents collaborate to identify the correct elements on the page for navigation and data retrieval. They establish links with HTML code elements and then plan and translate actions into Python code for execution. This streamlines web data extraction, enabling developers to efficiently gather critical information for research, analysis, and application development. This automation significantly reduces manual coding efforts and improves the accuracy and speed of data acquisition.
- **Contract management**: Automating the extraction of key information from contracts using LLM, OCR, and VLM technologies. This solution enhances contract management, streamlines maintenance, and supports regulatory monitoring by tracking, standardizing, and organizing contract data for improved compliance and efficiency. This reduces the risk of missed deadlines, non-compliance, and inaccurate reporting by providing a centralized and automated system for managing contract details. This also improves efficiency in legal and procurement processes and strengthens regulatory oversight.

Instead of following pre-defined paths,

> "AI agents have planning abilities to craft their own journey and answer specific instructions."

Although these agent-based systems represent a significant breakthrough in innovation, like all AI systems, they have their limitations. Their use of third-party products for specific tasks significantly reduces hallucinations but does not eradicate them. Agent-based processes are particularly effective for verifiable problems, where an operator can validate the information retrieved by the agents. However, more complex tasks still face challenges due to a lack of trust in the system's outputs.

Navigating GenAI Limitations and Threats

Model Hallucination

The release of Google's Gemini 1.5 model in the beginning of 2024 sparked a wave of controversy due to significant biases[60]. For example, when prompted to depict German soldiers from World War II or the Founding Fathers of the United States, the AI generated images of Black or Asian individuals. Although this issue was later addressed, the incident highlighted the potential failures of AI models, particularly when handling historically and culturally sensitive topics.

(59) Discover our SiaGPT solution and our AI Agents use case library at sia-partners.com
(60) Anand, N. (2024, February 22). Google's Gemini AI accused of acting too "woke", company admits mistake.

GenAI & cybersecurity threats

The ability of generative models to produce malicious code has raised significant concerns, making cybercrime more accessible. A bad actor, for instance, could prompt a chatbot to generate code for an SQL injection attack or quickly craft highly convincing phishing emails.

On the other hand, recent studies have shown that generative models may not necessarily enhance the capabilities of cyber attackers as expected. For instance, a study described in the Llama 3 herd of models technical report[61] measures the effectiveness of LLM assistance in cyberattack scenarios and finds that both novice and expert attackers demonstrated insignificant uplift in performance when using Llama 3 405B, compared to having open internet access without LLMs. Despite the advanced capabilities of the model, the challenge completion rates between the two cohorts were nearly identical, suggesting that generative models may not drastically improve the success rates for cyber attackers in real-world conditions.

Moreover, models can exhibit unintended behavior when manipulated with carefully crafted prompts. This exploit, known as jailbreaking, remains a persistent challenge in LLM security. While providers continuously patch vulnerabilities, new jailbreak techniques continue to emerge, highlighting the ongoing arms race between security researchers and adversaries.

Another major concern is data resurgence, where models unintentionally reveal proprietary or confidential information. As many users input sensitive data into tools like ChatGPT[62], models trained on publicly available or user-submitted data may inadvertently expose it.

A striking example involved an LLM reproducing Samsung's internal documentation when fed a partial prompt, demonstrating the potential risks associated with GenAI models handling proprietary data.

In response to these challenges, recent progress has been made in the area of model safety. In particular, the activation of ASL 3 in the Claude 4 Opus release[63] represents a key step in improving the security of generative AI. ASL 3 incorporates enhancements such as the ability to refuse to respond to requests that violate security policies, including forbidden content and attempts to extract private information. These improvements guide model behavior more effectively towards safer and more reliable results.

ANSSI's[64] guidelines emphasize adopting a lifecycle approach to security, integrating governance frameworks, and leveraging collaboration to stay ahead of evolving threats. These include:

- Implementing security by design throughout the AI lifecycle with proper governance frameworks and human validation
- Integrating AI systems into existing security operations despite forensic challenges
- Developing tailored governance policies to address the gap between security policies adoption (majority of companies) and actual defense capabilities
- Establishing AI red teams and partnerships with specialists to validate trust levels
- Conducting employee training to prevent misuse and data breaches when using GenAI tools.

(61) The Llama 3 herd of models. (n.d.). Retrieved from https://ar5iv.labs.arxiv.org/html/2407.21783
(62) Writer, R. L. C. (2023, December 8). Employees are feeding sensitive Biz data to ChatGPT, raising security fears.
(63) Anthropic activating ASL 3: Enhancing Safety and Alignment in Claude 4 Opus: https://www.anthropic.com/activating-asl3-report
(64) The French National Cybersecurity Agency: Agence nationale de la sécurité des systèmes d'information (ANSSI)

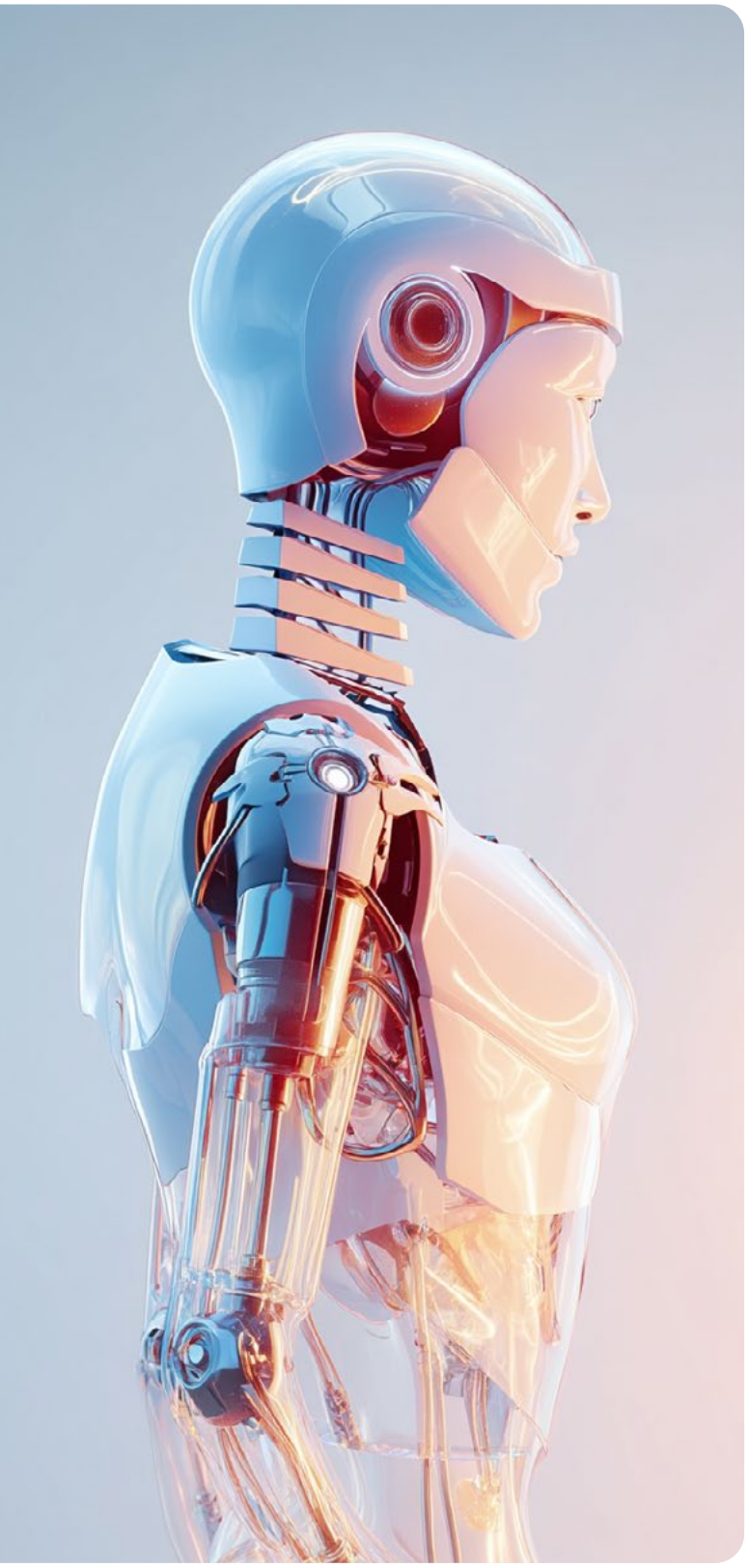
Models misuse, and its threat to democracy

Another growing concern is the use of deepfakes. In the past years, several scandals emerged involving defamation through degrading AI-generated images of real individuals, ranging from celebrities to high school students. Deepfakes also pose a serious threat to democracy due to their potential for disinformation. This was evident in the recent Indian elections[65], where fabricated posts falsely depicted deceased politicians.

The increasing accessibility of advanced generative models (both open-source, such as Stable Diffusion 3.5, and proprietary, e.g. OpenAI's SORA or X's Grok-2) has made it easier than ever to create high-quality malicious content. Disinformation campaigns leveraging AI have become more prevalent, with Russia, for example, using fake accounts and AI-generated images for political destabilization . While there is still limited evidence that these tactics have had a significant impact, they are being closely monitored.

In response, initiatives have emerged to tackle these challenges, including an annual artificial intelligence summit jointly organized by the United Kingdom and South Korea, leading to the establishment of a network of AI safety institutes and a commitment by leading AI companies to voluntary AI safety standards[67].

(65) Landrin, S. (2024). India's general election is being impacted by Deepfakes.
(66) Iyengar, R. (2024, November 4). Russia behind fake Haitian voter election videos, U.S. officials say. Foreign Policy. Retrieved from https://foreignpolicy.com
(67) South Korea's Tech Prowess Takes Centre Stage at AI Seoul Summit



Anticipating The Future: Preparing For 2026 and Beyond

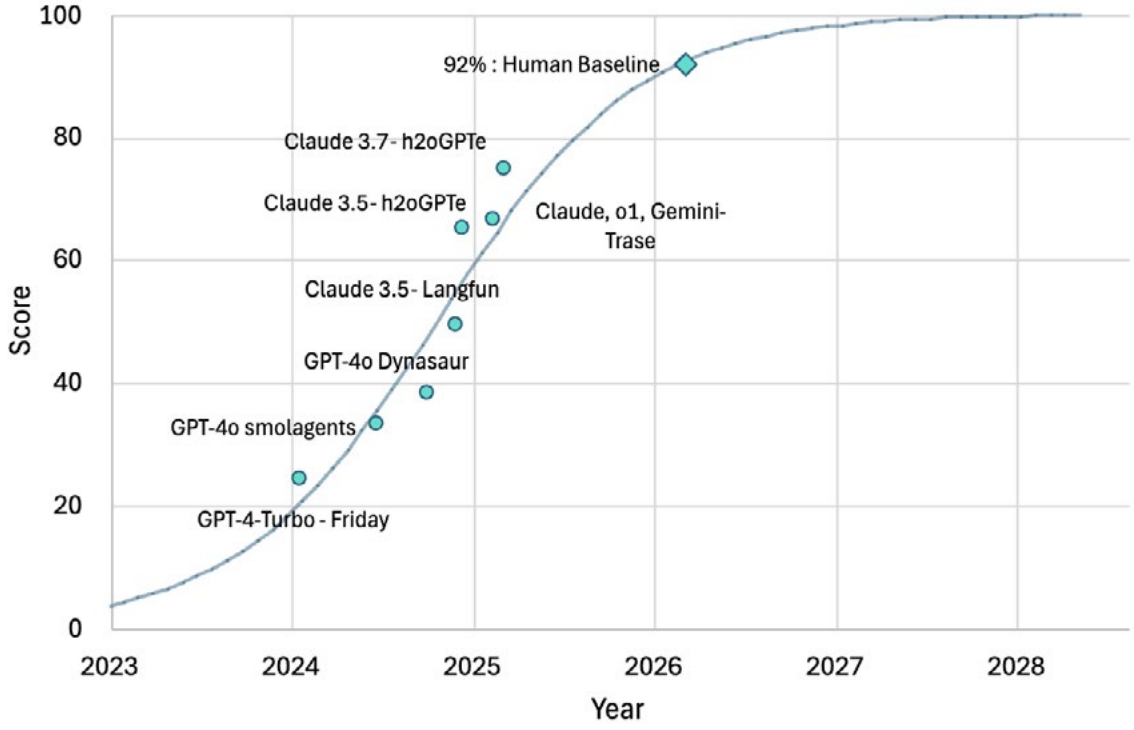
5. Anticipating the future: Preparing for 2026 and beyond

As the pace of AI innovation accelerates, anticipating future trends is critical for businesses preparing for the years ahead. This section looks towards 2026 and beyond, exploring key developments in AI agents and their integration into robotics that promise to redefine what's possible. We will dive into the perspectives for research and development, the evolution of AI evaluation, and the exciting applications of AI in fundamental research and physical systems, providing insights into the technologies and trends that will shape the future of work and innovation.

Agentic Development will be the hottest topic

FIGURE 22: AI AGENTS ARE QUICKLY PASSING THE S-CURVE OF CAPABILITY
Source: Aymeric Roucher (Hugging Face)

AI agents are quickly passing the S-curve of capability
AI are becoming more and more capable duo to the development of multimodal capabilities, the growth of Agentic Frameworks to enable their creation, and an ever increasing suite of tools and protocols that Agents can use. The score is the percentage of correct answers on the GAIA **(General AI Assistants)** benchmark's test set.



The last couple of years have seen remarkable advancements in AI agents, demonstrating not only their feasibility but also their high relevance in a variety of real-world contexts, such as ERP Automation, HR Automation, AI Code Review or Information Retrieval. This trend is set to accelerate significantly in the coming years, driven by four several key factors:

01 **The Continued Development of Multimodal Models:** These models are capable of processing and generating multiple types of inputs and outputs, adapting to prompts, and functioning as reasoners when needed. They will simplify user interactions and expand the possibilities for engaging with tools, systems, and applications. For instance, the newly released Gemini 2.0 can process and generate text, images, audio, and video. This flexibility enables a wide range of applications, including image analysis, enhanced OCR (Optical character recognition), and efficient user interactions across different platforms. For instance, on Google TV, users can ask questions about movies or general knowledge and receive relevant content from YouTube[68].

(68) Krol, J., & Schwanke, A. (2025). The next-generation of Google TV is on the way with an improved Gemini that'll make smarter and better.

02 **The Growth of Frameworks for Seamless AI Agent Creation:** Frameworks that simplify the creation of AI agents are becoming increasingly popular. A prime example is Hugging Face's SmolAgents, which quickly gained recognition with over 3,000 stars on GitHub within months of its release. This framework streamlines the agent creation process by providing essential tools such as search capabilities and dynamic execution.

Nvidia's recent introduction of **agentic AI blueprints** at CES 2025 further underscores this momentum. These blueprints are designed to help developers build AI agents capable of solving complex tasks, like searching and summarizing videos[69]. This underscores the growing significance of agent-based AI. This release at one of the most significant annual tech events highlights the increasing importance of agentic AI.

03 **The improvements in high performance models:** Frameworks that simplify the creation of AI agents are becoming increasingly popular. A prime example is Hugging Face's SmolAgents, which quickly gained recognition with over 3,000 stars on GitHub within months of its release. This framework streamlines the agent creation process by providing essential tools such as search capabilities and dynamic execution.

Small, fast, and high-performing models can efficiently route and break down queries, while larger flagship models can be reserved for the most challenging tasks.

04 **The growing capacity of AI models to use external tools:** AI models are increasingly capable of interacting with tools in sophisticated ways. What was once limited to automating simple tasks, such as drafting emails or preprocessing invoices, has evolved into AI agents making fully autonomous decisions in complex environments. For example, Google DeepMind utilizes AI agents for smart grid management, where they autonomously balance competing goals such as energy efficiency and equipment longevity, resulting in significant reductions in power consumption[70].

In particular, the ability to **use digital tools along with multimodal image and video processing** will enable models to become better assistants and expand their interactions with users.

New benchmarks will emerge for AI Performance and Reasoning Eval

As AI model performance rapidly advances, the benchmarks used to evaluate them are also evolving. Traditional benchmarks like MMLU (for general tasks), HumanEval (for coding), and GPQA (for PhD-level questions) are becoming saturated, as leading models approach 90% accuracy and performance differences narrow[71 72]. Furthermore, there's a growing risk (likely true) that existing benchmark data has leaked into training datasets, artificially inflating performance scores. This necessitates the development of new standards and criteria for evaluating AI capabilities.

A key focus of these new benchmarks is on evaluating reasoning abilities. To better assess reasoning, benchmarks like **ARC-AGI**[73] (measuring skill acquisition and unfamiliar task solving) and **GPQA Diamond**[74] (challenging models with PhD-level science questions) are gaining prominence. These tests were prominently featured during the demonstration of OpenAI's GPT-o3 capabilities at the 12 Days of OpenAI event[75].

(69) Boitano, J. (2025). Nvidia and partners launch agentic AI blueprints to automate work for every enterprise.
(70) Top 20 Agentic AI Use Cases in the Real World
(71) LLM Benchmarks: Overview, Limits and Model Comparison
(72) Chatbot Arena (formerly LMSYS): Free AI Chat to Compare & Test Best AI Chatbots
(73) What is Arc-Agi? (n.d.)
(74) Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., ... Bowman, S. R. (2023). GPQA: A graduate-level google-proof Q&A benchmark.
(75) OpenAI o3 breakthrough high score on Arc-AGI-pub. (n.d.-a).

The rise of agentic models, capable of autonomously performing complex tasks, adds another layer of complexity to the evaluation. Accurately assessing the agentic capabilities of these models is now essential. **METR (Model Evaluation & Threat Research)**, a nonprofit founded in 2023, is dedicated to evaluating the autonomous capabilities of advanced AI systems and their potential risks. **METR** provides Autonomous Evaluation Resources, designed to assess an AI's ability to perform complex, multi-hour tasks without human intervention. These resources have already been used to evaluate several models, and their adoption is expected to become increasingly widespread.

These developments have several key implications for businesses:

Shifting Evaluation standards:

Traditional benchmarks are becoming less effective at differentiating AI models, making it critical for businesses to stay updated on new evaluation standards.

Internalizing Benchmarking Capabilities

Given the risk of benchmark data leaks, organizations should develop and internalize their own robust AI evaluation capabilities. They also need to prioritize transparency and rigor in their AI evaluation processes.

Competitive Advantage:

Performing well on new, tougher benchmarks offers a significant competitive edge for AI models.

AI Risk Assessment:

As autonomous AI models become more prevalent, businesses must prioritize safety and ethical considerations. Even minor input errors can lead to incorrect decisions, which, when amplified across multiple autonomous agents, may create cascading failures with severe consequences. For example, a slight misinterpretation by an autonomous vehicle could escalate into a critical safety error, leading to an accident.

Even as benchmarks evolve to better evaluate the growing capabilities and performance of GenAI models, limitations may still emerge in other areas, such as data quality, business logic, or deployment constraints, which can reduce performance in specific use cases.

A Broader Scientific Perspective

AI for fundamental research

In the broader scientific and technical landscape, AI is making substantial contributions beyond classical areas such as NLP or image processing. A MIT study[76] on AI's impact on scientific discovery revealed that

> "AI-assisted researchers discovered 44% more materials, and that it led to a 39% increase in patent filings[77]"

AI also increased downstream innovation by +17% according to that same study.

(76) FunSearch: Making new discoveries in mathematical sciences using Large Language Models
(77) Aidan Toner-Rodgers (2024). Artificial Intelligence, Scientific Discovery, and Product Innovation.

These statistics aren't merely academic; they signify a fundamental acceleration in the pace and potential of innovation across industries. LLMs empower individuals to better leverage unstructured information and access advanced human intelligence, paving the way for groundbreaking achievements across diverse fields:

Drug Discovery:

AI, especially deep learning models trained on vast datasets, can accelerate the identification of promising drug candidates by navigating the complex landscape of possible chemical configurations.

Materials Science:

AI can generate "recipes" for novel compounds with predicted properties, potentially revolutionizing materials design.

Structural Biology:

In this field, where fundamental principles are known but complex rules make identifying specific instances challenging, AI assistants can significantly accelerate discovery.

Genomics:

Similar to structural biology, genomics involves analyzing complex data, an area where AI excels at identifying patterns and insights.

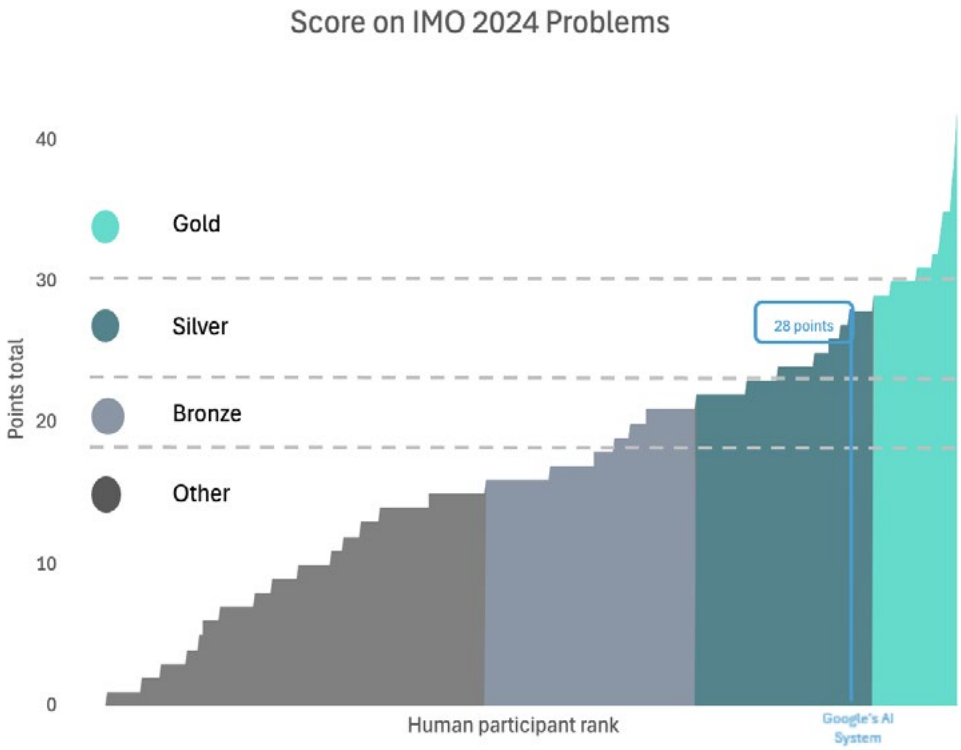
Climatology:

AI can analyze vast and complex climate datasets, potentially accelerating discoveries and informing more effective climate action strategies.

Meanwhile, advanced models such as AlphaProof and AlphaGeometry 2 are pushing the boundaries of AI-assisted reasoning. These systems recently achieved a silver medal performance at the International Mathematical Olympiad[78] (IMO) (see **Figure 23**).

Together, these advancements represent a new paradigm in scientific discovery, demonstrating how AI can complement human intuition to unlock deeper understanding.

FIGURE 23: GOOGLE DEEPMIND'S AI SYSTEMS IMO 2024 SCORE
Source: Deepmind's AlphaGeometry & AlphaProof teams



AI for robotics

On-device Generative AI (GenAI) is poised to accelerate the development of robotics applications, largely due to advancements in multimodal models. These models expand the scope of robotics by enabling robots to autonomously process and respond to diverse inputs, including audio, images, and video, and generate varied outputs, such as movements, text, and audio.

> This enhanced multimodal capability moves robotics closer to achieving human-like interaction and functionality, paving the way for "Physical AI."

AI also increased downstream innovation by +17% according to that same study.

Google DeepMind's AutoRT[79] system already utilizes Vision-Language Models (VLMs) to interpret a robot's environment, translate it into descriptive language, and then leverage a Large Language Model (LLM) to process this information and dynamically suggest sequential tasks[80]. As with AI Agents, the model encodes these tasks as prompts and converts them into physical action commands.

As with LLMs, open source is paving the way for increasingly accessible robotics, thanks to open frameworks and tools like Open Robotics[81] and Hugging Face's LeRobot[82]., which provides pre-trained models, datasets with human-collected

(78) AI achieves silver-medal standard solving International Mathematical Olympiad problems
(79) AutoRT: Embodied Foundation Models for Large Scale Orchestration of Robotic Agents
(80) Shaping the future of advanced robotics. (2024, December 17).
(81) Open Robotics
(82) HuggingFace's Le Robot

demonstrations, and pre-trained examples designed for real-world robotics applications.

Nvidia's video model, Cosmos[83], generates synthetic data for training robots in diverse environments[84], enhancing their adaptability and performance. This capability addresses a key challenge in robotics: the need for vast and varied training data sets.

Open-source projects, such as OpenVLA (Visual-Language-Action model), integrate computer vision and natural language processing to enable agents to perceive their environment, understand instructions, and perform actions. These developments will expand AI use cases to include physical applications:

01 **Manufacturing and Industrial Automation:**

- **Advanced Assembly Processes:** AI-equipped robots can assemble parts with exceptional precision and dynamically adjust movements based on real-time data analysis, significantly improving production speed and product quality.
- **Smart Factories:** Physical AI facilitates the creation of autonomous factories where robots handle a wide range of tasks, from assembly and quality control to maintenance and logistics, with minimal human oversight.

02 **Warehouse Automation with Autonomous Mobile Robots (AMRs):** AMRs automate the transportation of goods within warehouses and across outdoor settings, reducing reliance on manual labor and dramatically increasing operational efficiency.

03 **Healthcare and Assistive Robotics:**

- **Surgical Assistance:** Robots enhance surgical precision, assist in complex procedures, and minimize invasiveness, leading to improved patient outcomes.
- **Rehabilitation and Elderly Care:** While never replacing human interaction, AI-driven robots can provide crucial support to carers, assisting with physically demanding tasks like lifting and manipulation, thereby enhancing the quality of care and safety for both caregivers and patients.

04 **Agriculture and Precision Farming:** Physical AI enables the further optimization of crop management and resource consumption, leading to increased yields, reduced environmental impact, and more sustainable agricultural practices.

05 **Autonomous Vehicles and Simulation-Based Training:** Platforms such as NVIDIA Cosmos enhance the safety and decision-making capabilities of autonomous vehicles through realistic simulated training environments, accelerating their development and deployment.

06 **Adaptive Robotics:** Robots can operate in dangerous or hard-to-reach environments, providing critical support in disaster response, environmental remediation, and other high-risk situations.

(83) Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., & Cai, T. (2025). Cosmos world foundation model platform for physical ai.
(84) Liu, M. (2025, January 11). Cosmos World Foundation models openly available to physical AI developers | NVIDIA blog.

Taking Action: Concrete Steps for Leaders

6. Taking Action: Concrete Steps for Leaders

Generative AI is reshaping work across industries, forcing to rethink people, process and technology together. Rather than just deploying tools, organizations need a strategy that aligns workforce transformation, data practices and infrastructure with business goals. This section provides practical guidance on navigating this transformation. We will explore how to reshape workforce management for the AI era, address new ways of working with data and integrated AI solutions, evolve infrastructure to support context retrieval, and navigate the decision between cloud and on-premise model deployment, all while considering the evolving roles and skill sets needed for the future.

Human Resources for AI: Reshaping Workforce Management for the AI Era

AI is redefining businesses and workforce management, requiring new skills, roles, and strategies. HR must:

01 **Integrate AI into its own functions** (e.g., talent acquisition, benefits analysis).

02 **Address the risk of AI-assisted cheating during interviews** (e.g., technical tests, limited video calls).

03 **Prepare the workforce for an AI-driven future with AI-tools trainings.**

According to Gartner's Reskilling and Upskilling the Workforce report[85], generative AI would require 80% of the engineering workforce to upskill[86] redefining the boundaries between junior and senior roles. As AI takes over certain tasks, expectations for each level will shift fundamentally.

This means that HR will be at the forefront of guiding organizations through change, by making us rethink what we expect of the workforce of the future.

From a strategic standpoint, HR team should work closely with top management to build a clear GenAI vision aligned with the company's mission, emphasizing not only digital literacy but also strategic thinking, adaptability, and effective collaboration with AI systems. From an operational perspective, preparing the workforce involves implementing robust upskilling programs, embedding GenAI tools to support daily tasks, fostering knowledge sharing, and clearly communicating the dos and don'ts of using GenAI within the organization. To ensure long-term sustainability, these changes must be seamlessly integrated into HR processes - from recruitment and onboarding to career development - enabling a workforce that is both AI-augmented and future-ready.

Although GenAI will impact all types of workforces, one of the most striking examples of its effect is the future of software development. Generative AI is reducing the need for basic coding skills, and therefore the demand for average developers. Instead, organizations will seek excellence: only the best, those combining deep technical expertise, critical thinking, and business acumen, will be essential to harness this new ecosystem, especially in managing the growing volume of AI-generated code.

How will generative AI make average developers obsolete and great developers more valuable than ever?

01 **Automation of basic tasks:** Repetitive, low-complexity work is now handled by AI, reducing the value of pure operational skills.

02 **Need for strong human expertise**

- Although AI is effective, it may generate code that contains errors or is suboptimal, particularly when dealing with complex problems[87].
- Experienced developers are key to spotting and fixing these flaws, ensuring quality in both technical and business terms[88].

(85) Reskilling the Workforce
(86) Gartner Says Generative AI will Require 80% of Engineering Workforce to Upskill Through 2027
(87) Mckinsey : Unleashing developer productivity with generative AI
(88) Deloitte : How can organizations engineer quality software in the age of generative AI?

03 Redefinition of the developer's role:

- Development is evolving into rapid prototyping, supported by generative AI[89].
- Senior developers increasingly function as AI orchestrators: guiding, supervising, and refining AI code contributions.
- Advanced skills in software architecture, performance optimization, and strategic decision-making are more critical than ever.
- Developers also increasingly act as proxy Product Owners, aligning technical solutions with business needs and easing the burden on product teams[90].

AI Adoption and Business Impact

Leading companies are moving beyond isolated AI pilots and toward enterprise scale deployment, recognizing that broad adoption, rather than narrow experimentation, drives tangible returns. Economists note that higher rates of AI adoption can significantly boost productivity and improve quality outcomes[91]. Recent OECD analysis estimates that AI could raise annual labour productivity growth by between 0.4 and 1.3 percentage points in highly AI-exposed economies such as the United States and the United Kingdom, and by 0.2 to 0.8 points in others like Italy and Japan[92]. This transition demands more than just technology, it requires strategic alignment across culture, organization and execution.

Managers often struggle to understand how AI addresses real business problems and may underestimate the organizational and cultural shifts required.

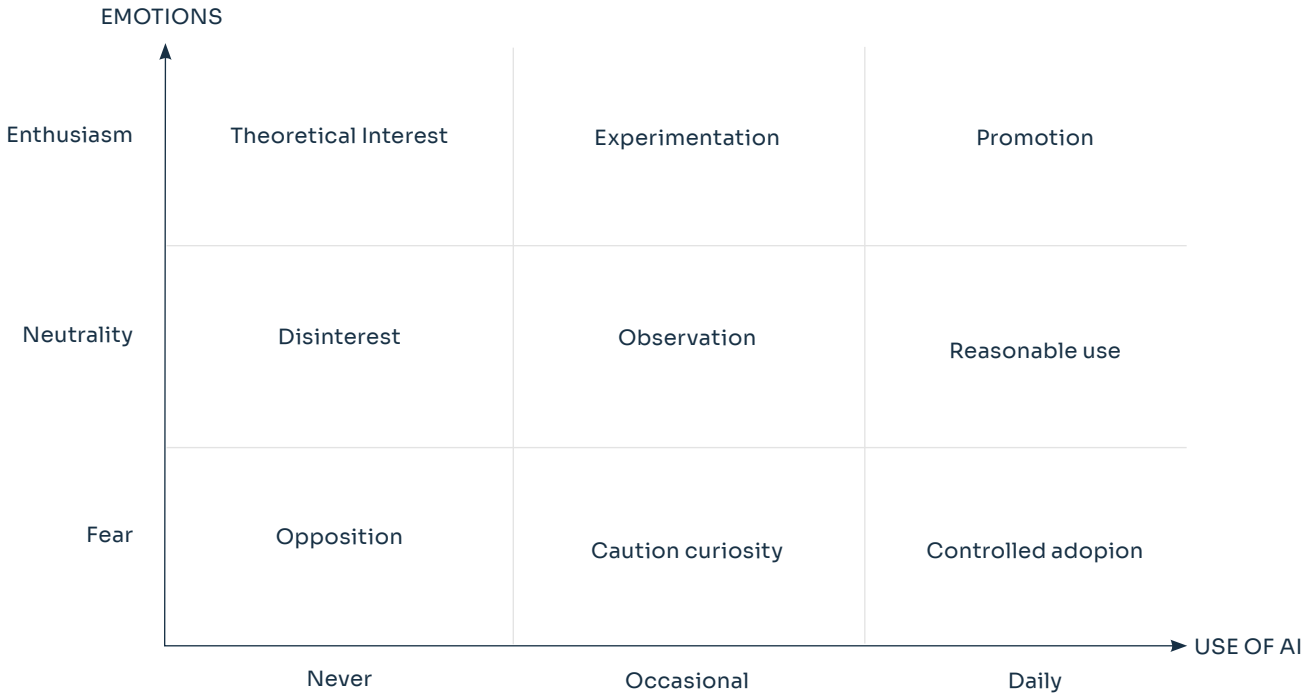
Successful firms therefore invest in comprehensive training, clear communication and change management to build fluency in AI throughout the workforce. Without such alignment, even the most promising technologies may fail to take root.

Executive leadership plays a pivotal role in driving AI fluency and organizational changes. Senior leaders must articulate a clear vision for AI integration, prioritize investments in skills and tools"future-ready" with a clear AI strategy[93], highlighting the urgent need for leadership to drive AI adoption and build workforce capabilities that keep pace with AI-driven transformation.

Understanding employee mindsets is key to tailored adoption strategies. For example, a framework developed by ESSEC and Sia Partners (**Figure 24**) plots employees in an adoption and emotion matrix based on their engagement and sentiment. This approach identifies segments such as enthusiastic experimenters, cautious adopters and disengaged skeptics. Leaders can then target each group appropriately: curious experimenters may be invited to pilot projects, while more hesitant teams receive focused training and reassurance. In this way, change management can be calibrated to the specific attitudes and readiness of each group.

(89) PWC : 10 ways GenAI improves software development
(90) Mckinsey : How an AI-enabled software product development life cycle will fuel innovation
(91) The Adoption of Artificial Intelligence in Firms
(92) OCDE : Macroeconomic productivity gains from Artificial Intelligence in G7 economies
(93) Adecco Group : Only 10% of C-suite leaders say their companies are ready for AI disruption, finds latest Adecco Group report

/FIGURE 24: ESSEC-SIA PARTNERS FRAMEWORK: EMPLOYEE ADOPTION AND EMOTION MATRIX



Closing the strategy to execution gap is essential for realizing AI's business impact. AI initiatives should be tied to clear metrics and business objectives, with ongoing feedback loops to monitor progress and adapt plans.

By regularly reviewing outcomes and iterating based on results, organizations ensure that strategy remains grounded in operational reality. This continuous loop between planning and implementation helps maintain momentum and delivers on the promise of AI.

New ways of working: Data, Infrastructure & Integrated AI solutions

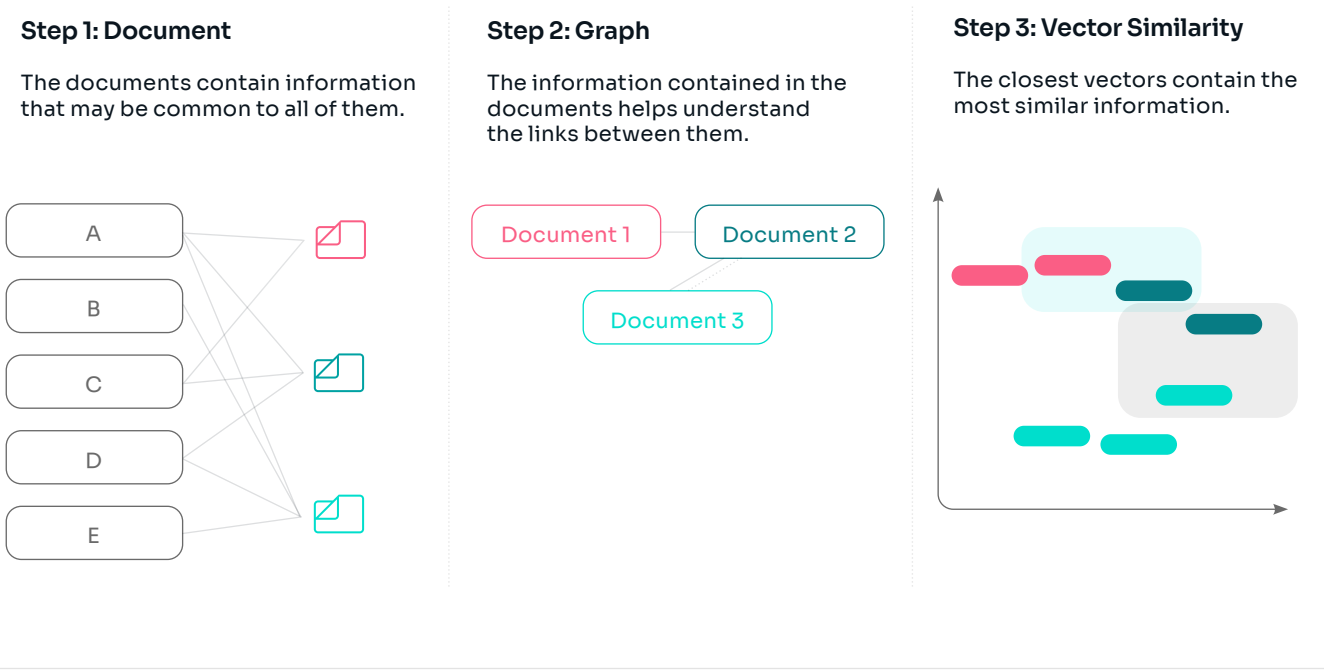
Infrastructure Evolution: Shifting to Context Retrieval

The rise of GenAI represents a significant shift in how organizations manage their technology infrastructure and workforce. Traditional data science often relies on skills such as exploratory data analysis (EDA), feature engineering, model training, and hyperparameter tuning.

However, generative AI challenges this approach by emphasizing the importance of contextual information over structured datasets or pattern recognition from unsupervised learning.

This shift requires organizations to rethink their workflows, moving away from data labeling and training to building systems that can efficiently retrieve and apply relevant context to AI models. Solutions such as vector databases, which enable semantic search and fast retrieval of relevant information, are essential for generating accurate and context-aware outputs from generative AI models.

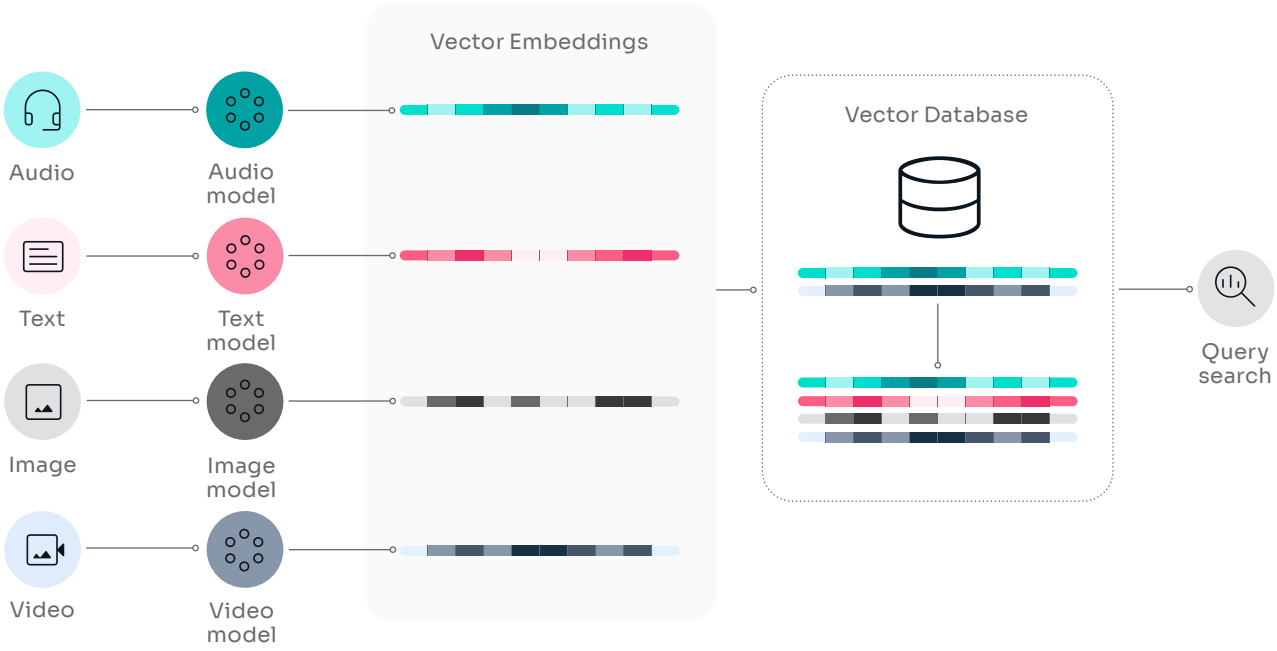
FIGURE 25: VECTOR SIMILARITY IN A GRAPH CONTEXT



Solutions such as vector databases (**Figure 25, Figure 26**), which enable semantic search and fast retrieval of relevant information, are essential for generating accurate and context-aware outputs from generative AI models. As generative AI becomes more integrated into complex processes, managing post-inference outputs, integrating with external tools, and orchestrating sophisticated chains of thought (CoT) are increasingly important.

This transformation calls for not only a change in how data is accessed but also in how AI models interact with dynamic, evolving contexts, as we discuss in the section "Retrieval Augmented Generation (RAG): The Indispensable GenAI Use Case for Businesses."

/FIGURE 26: VECTOR EMBEDDINGS & VECTOR DATABASES CONCEPT

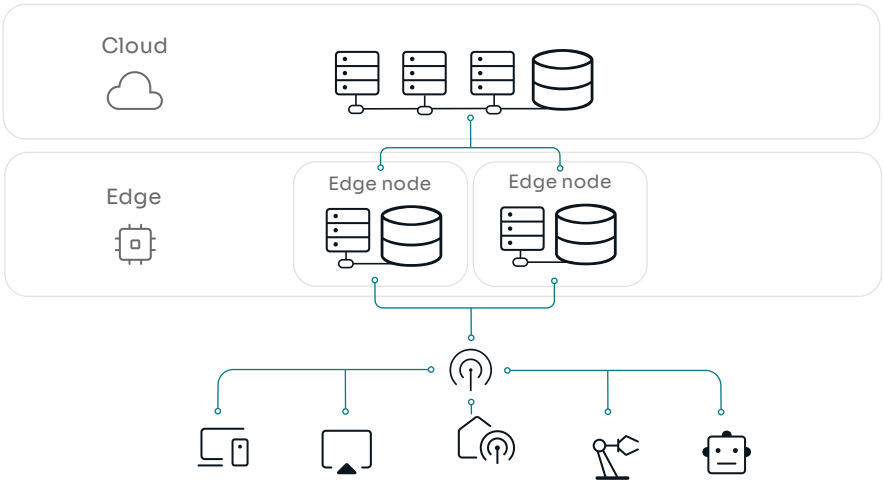


Model Deployment: Cloud vs. On-Premise Solutions

Organizations must assess deployment options based on their specific needs and privacy considerations. The decision between cloud-based, on-premise and edge computing solutions has significant implications for factors such as budget, performance, and security:

- Cloud-based Solutions: Cloud platforms offer flexibility, scalability, and faster deployment times. These solutions are often preferred for organizations that require quick integration, easy management, and the ability to scale their operations with minimal infrastructure investment. They also provide access to advanced AI tools and computing resources, making it easier to develop and run sophisticated generative models without the need for heavy internal infrastructure. By integrating Google Cloud's Gemini with services such as BigQuery for data analysis and Vertex AI for model deployment, companies can efficiently harness AI-driven insights to personalize customer experiences, predict trends, and scale their operations, all within a unified ecosystem.
- On-Premise Solutions: On the other hand, some organizations, especially those with stringent data confidentiality or compliance requirements, may prefer on-premise hosting. This ensures that sensitive data never leaves the organization's network, providing an additional layer of control over data security and privacy. However, on-premise deployments come with increased setup and maintenance costs, as well as the need for specialized internal expertise to manage and update infrastructure.
- Edge Computing Solutions: For environments where workloads need to run even closer to where data is generated - such as in industrial settings, IoT applications, or remote locations - edge computing provides a interesting alternative. By processing data locally on edge devices or nearby gateways, organizations can reduce latency, improve real-time responsiveness, and ensure continued operation even when connectivity to centralized cloud infrastructure is limited or unreliable. This approach also supports data sovereignty (**Figure 27**) and minimizes bandwidth usage, making it relevant for applications involving real-time monitoring, predictive maintenance, or autonomous systems operating in the field.

/FIGURE 27: EDGE COMPUTING ARCHITECTURE



New Workflows & Evolving Roles

The rise of generative AI and its associated technical complexity has led to the emergence of specialized roles in organizations, such as "GenAI Engineers". These professionals combine software engineering skills, AI implementation expertise, and a strong understanding of cloud technologies. Their hybrid skill set emphasizes practical deployment and integration over theoretical machine learning knowledge, reflecting a broader industry trend toward leveraging pre-trained models rather than developing them from scratch. Organizations that invest in cultivating this talent will gain a significant edge in their AI transformation, enabling them to deploy solutions more efficiently and at scale.

At the same time, generative AI is reshaping the workflows of non-technical professionals, extending their capabilities and boosting their productivity. Tools such as ChatGPT, Copilot and other industry-specific AI tools streamline day-to-day tasks such as content writing, while enabling them to tackle challenges traditionally handled by technical teams, such as minor code tweaks or data analysis. By democratizing access to these capabilities, these tools enable non-technical teams to contribute to the organization's objectives in a new way.

To support this evolution, organizations need to focus on providing intuitive tools, integrating them seamlessly into workflows.

Additionally, they should implement targeted training programs to optimize use and efficiency and put in place robust data governance to ensure trust and foster widespread adoption across all roles.

Ultimately, this dual evolution underscores a critical insight: Generative AI, at its core, is rapidly becoming a specialized domain within software engineering. As models grow more advanced and ubiquitous, the demand for sophisticated software skills to design, build, and run robust GenAI applications will only intensify. Organizations that recognize and act on this reality, by investing in both specialized engineering talent and the seamless integration of AI tools, will be best positioned to harness the full transformative power of GenAI.

Conclusion

Over the past couple of years, there has been an explosion in applications of generative AI, driven by major technological advances and widespread industry adoption. As AI systems become more capable, powerful, and embedded across sectors, conversations about their ethical use, security, and global impact have gained ground. This major shift in the way we work has opened the door to new opportunities while also presenting significant challenges and obstacles that will shape the future of AI development and governance.

Technologically, models have become not only more powerful but also more versatile, thanks to breakthroughs in multimodality and techniques like retrieval-augmented generation (RAG), making LLMs more relevant for organizations. This increased adaptability has enabled GenAI to make deeper inroads into sectors ranging from corporate innovation to fundamental research.

At the same time, GenAI applications have shifted, transforming from a research-driven innovation into a core pillar of economic strategy for corporations. GenAI companies are increasingly prioritizing real-world applications over foundational model development, driving record-breaking revenue growth. This transformation is further fueled by the rise of on-device AI, which enhances privacy, efficiency, and accessibility. These trends are set to accelerate in the coming years.

However, this rapid progress in adoption and performance also brings heightened risks. The rise of deepfakes, cyber threats, and the growing dominance of a few AI giants have intensified concerns around safety, ethics, and monopolistic behavior. As a result, governments now see generative AI as a critical geostrategic issue, with some tightening regulations to mitigate risks while others focus on securing a strategic edge in an increasingly AI-driven global economy.

Looking ahead, we anticipate the rise of agentic models capable of autonomous reasoning and decision-making, which will necessitate new testing methods and safety guardrails. The expansion of multimodal AI will unlock groundbreaking possibilities, but it will also introduce new regulatory challenges.

The AI revolution is poised to reshape society at least as profoundly as the industrial and digital revolutions, if not more. AI models will soon assist us in most aspects of daily life, much like machines, engines, the internet, and smartphones already do. In many ways, LLMs represent humanity's collective knowledge, culture, and values, compressed into intelligent assistants. It is therefore crucial to ensure the development of diverse AI models that reflect the unique languages, cultures, and perspectives of different societies worldwide. For those reasons, to lead effectively in this era of rapid and complex GenAI evolution, it is imperative to first understand the critical dynamics of this shifting landscape to make informed, strategic decisions tailored to your unique context.

Sia stands ready to guide your organization through this transformative journey. Contact us today to discuss your specific AI project needs and explore how our expertise can help you navigate the complexities and capitalize on the immense opportunities of generative AI.

/ Steering Committee

Alexandre ORHAN
Senior Manager, R&D Lead
AI & Data
alexandre.orhan@sia-partners.com

Théophile LOISEAU
Manager, Lead GenAI
AI & Tech Foundations
théophile.loiseau@sia-partners.com

Arnaud TATIN
Managing Director
AI & Tech Foundation
arnaud.tatin@sia-partners.com

Alexandre LALAU
Associate Manager
AI & Data
alexandre.lalau@sia-partners.com

Kaushik MOUDGALYA
Senior Consultant
AI & Data
kaushik.moudgalya@sia-partners.com

Yasmine BENNANI
Senior Consultant
AI & Data
yasmine.bennani@sia-partners.com

Axel DARMOUNI
Senior Consultant
AI & Tech Foundations
axel.darmouni@sia-partners.com

Ariel GUIDI
Senior Consultant
AI & Tech Foundations
ariel.guidi@sia-partners.com

Victor DE CHAISEMARTIN
Consultant
AI & Data
victor.dechaisemartin@sia-partners.com

AI Model Benchmarks Primer

AIME

American Invitational Mathematics Examination, a prestigious **high-school math competition** with 15 increasingly difficult problems. The median score is **4–6 correct** answers out of 15. It involves questions of increasing difficulty, with the answer to every question being a single integer from 0 to 999.

ArenaHard

A **dynamic, real-world benchmark** from LMSYS Org that:

- **Differentiates model capabilities** effectively.
- **Reflects human preferences** in real-world applications.
- **Updates frequently** to prevent overfitting.

Uses **GPT-4-Turbo** for evaluation and features **500 challenging queries** from the Chatbot Arena.

Codeforces

A **competition-level coding benchmark** using real-time submissions on Codeforces to assess LLMs with human-comparable **ELO ratings**. Ensures **zero false positives** and includes special judges for fairness.

FrontierMath

A benchmark of hundreds of original, expert-crafted mathematics problems designed to evaluate advanced reasoning capabilities in AI systems.

GAIA

GAIA evaluates reasoning, multimodal understanding, web browsing, and overall tool-use proficiency. Unlike traditional benchmarks that test highly specialized knowledge, GAIA presents conceptually simple questions that are easy for humans but remain challenging for even the most advanced AI models. For instance, human respondents achieve a 92% success rate, whereas GPT-4 with plugins scores only 15%. This contrast highlights a key gap in current AI capabilities, diverging from the trend of LLMs surpassing human performance in technical fields like law and chemistry. GAIA's approach suggests that true Artificial General Intelligence (AGI) will require robustness comparable to an average human's adaptability to everyday reasoning tasks. The benchmark consists of 466 curated questions, with answers to 300 withheld.

GPQA:

A benchmark featuring expert-level biology, physics, and chemistry questions. Even PhDs score only **65%** on average.

L

Livebench

a benchmark that aims to prevent LLMs from guessing answers which might as the related prompts / questions might have been part of the training dataset, since LLMs are trained on the "entirety of the internet". Livebench prevents such cheating by having questions based on recently released datasets, arXiv papers, news articles, and IMDb movie synopses)

LiveCodeBench

A **real-time, contamination-free** coding benchmark evaluating LLMs beyond code generation, including **self-repair, execution, and test prediction**. Collects problems from **LeetCode, AtCoder, and Codeforces**, ensuring continuous relevance.

M

MATH-500

A subset of **500 problems** from OpenAI's MATH benchmark, designed to assess advanced mathematical reasoning.

MMLU

The MMLU consists of about 16,000 multiple-choice questions spanning 57 academic subjects including mathematics, philosophy, law, and medicine. Designed to be more challenging than the last benchmark in a similar vein called General Language Understanding Evaluation (GLUE) on which new language models were achieving better-than-human accuracy.

MMLUPro

A harder than MMLU benchmark. **MMLU-Pro** improves upon the following aspect of MMLU:

- Expanding answer choices from **four to ten** for increased complexity.
- **Removing trivial and noisy questions** for a more rigorous evaluation.
- **Enhancing stability**, reducing prompt sensitivity from **4–5% to 2%**.

O

OpenVLM Leaderboard

This is a benchmark for the open source VLMs – Vision Language Models or API models that are publicly available. Hence, this is a benchmark for multimodal models.

S

SWE-Bench Verified

OpenAI's **human-validated subset** of SWE-Bench, offering a **more reliable** evaluation of AI models' ability to solve real-world software issues.

Terminology

A

Agentic AI / Agentic Development / Agentic Workflows

AI systems composed of multiple interacting models or components (agents) that can perform complex tasks autonomously. This includes planning, interacting with third-party products/tools, reviewing outputs, and setting new goals. It's a move beyond simply generating responses to actively solving problems and taking actions.

AI Act (European)

A comprehensive legal framework passed by the European Union in 2024 to regulate the development and use of AI. It's a key example of pro-regulation efforts.

AI Bubble

Concerns about inflated valuations of AI companies, particularly in comparison to their actual revenue and long-term profitability. The whitepaper raises this as a potential risk.

AI Safety Bill (SB 1047)

A proposed (and ultimately vetoed) California bill that aimed to regulate AI, including a controversial "kill switch" requirement.

Alignment (Model Alignment)

The process of ensuring that AI models' outputs and behaviors are consistent with human values, intentions, and ethical guidelines. A key concern in AI safety.

API (Application Programming Interface)

A set of rules and specifications that allows different software applications to communicate with each other. In the context of the whitepaper, it refers to the way businesses access and use AI models (e.g., GPT-4's API).

ARC-AGI (Abstract and Reasoning Corpus for Artificial General Intelligence):

A benchmark designed to measure the efficiency of AI skill-acquisition on unknown tasks.

C

Compute

Short for computational power. It is a crucial resource for training and running AI.

Context-Awareness

The ability of an AI model to understand and utilize the context of a situation or input to generate more relevant and accurate outputs. RAG is a key technique for enhancing context-awareness.

D

Data Contamination

The presence of AI-generated (synthetic) data within datasets used to train AI models. This can degrade model performance and is a growing concern.

Data Resurgence

The unintentional revelation of proprietary or confidential information by an AI model, often because the information was present in its training data.

Data Scarcity

Situation in which there is a shortage of readily accessible training data for AI model.

Deepfakes:

Realistic but fabricated media (images, videos, audio) created using AI, often used for malicious purposes like disinformation or defamation.

Deregulation (in AI)

The approach of minimizing government regulations on AI development and deployment, often advocated by tech companies to foster innovation.

FineWeb

A meticulously curated dataset developed by Hugging Face, used to demonstrate the positive impact of high-quality data on AI model performance.

Foundation Model

A large, general-purpose AI model (like GPT-4) that can be adapted to a wide range of tasks. The whitepaper discusses a shift from developing these to building applications on top of them.

GenAI (Generative AI)

A type of AI that can create new content, such as text, images, audio, and video, based on patterns learned from training data. This is the core focus of the whitepaper.

GPU (Graphics Processing Unit)

A specialized processor originally designed for graphics rendering but now widely used for AI training and inference due to its parallel processing capabilities. Nvidia is a dominant supplier.

HF (Human Feedback)

A crucial component often used in conjunction with Reinforcement Learning (RLHF - Reinforcement Learning from Human Feedback). It involves using human evaluations and preferences to guide the training of AI models. Humans provide feedback on the quality or appropriateness of model outputs, which is then used as a reward signal to improve the model's performance. This helps align the model with human values and desired behaviors.

Inference

The process of using a trained AI model to make predictions or generate outputs based on new input data. Generally uses a GPU for generation of these outputs.

Jailbreaking (LLM Security)

The process of crafting prompts to manipulate an AI model into bypassing its safety restrictions or generating unintended outputs.

L

LLM (Large Language Model)

A type of AI model trained on massive amounts of text data to generate human-like text, translate languages, write different kinds of creative content, and answer questions in an informative way. Examples include GPT-4, Llama, and Gemini.

LPU (Language Processing Unit)

A type of processor optimized for AI workloads, specifically designed for processing language-based tasks. Companies like Groq are developing LPUs.

M

MMLU (Massive Multitask Language Understanding)

A benchmark that evaluates a model's knowledge and reasoning abilities across diverse domains.

Mixture-of-Experts (MoE)

An AI model architecture that uses multiple "expert" sub-models, activating only the relevant ones for a given input. This improves efficiency and performance.

Model Quantization

A technique to reduce the size and computational requirements of AI models by representing their parameters with fewer bits. This improves efficiency and enables on-device deployment.

Mtokens

Millions of tokens.

Multimodal Models

AI models that can process and generate multiple types of data, such as text, images, audio, and video. This is a significant trend discussed in the whitepaper.

N

NLP (Natural Language Processing)

A field of AI focused on enabling computers to understand, interpret, and generate human language.

O

On-Device AI / On-Device GenAI

Running AI models directly on a user's device (e.g., smartphone, smart glasses) rather than relying on cloud servers. This enhances privacy, security, and responsiveness.

Open-Source Model

An AI model whose code and (sometimes) training data are publicly available, allowing others to use, modify, and distribute it. The whitepaper contrasts these with proprietary models.

P

Pro-Regulation (in AI):

The approach of advocating for stricter government regulations on AI development and deployment to address ethical concerns, safety risks, and potential misuse.

Proprietary Model

An AI model developed and owned by a specific company, often sold with a proprietary license, with restricted access to its code and its training data.

R

RAG (Retrieval-Augmented Generation)

A technique that combines the strengths of LLMs with information retrieval systems. It allows LLMs to access and incorporate relevant information from external sources (e.g., a company's internal documents) to generate more accurate and context-aware outputs.

Reasoning (in AI)

The ability of an AI model to go beyond pattern recognition and apply logic, inference, and problem-solving skills to arrive at conclusions or generate outputs. This is a key area of advancement in AI.

RL (Reinforcement Learning):

A type of machine learning where an AI agent learns to make decisions by interacting with an environment and receiving rewards or penalties for its actions. The whitepaper mentions RL in the context of post-training scaling laws, suggesting it's used to fine-tune models after their initial training. It's distinct from the initial training on large datasets.

S

SaaS (Software as a Service):

A software distribution model where applications are hosted by a provider and accessed by users over the internet. The whitepaper compares the revenue growth of GenAI companies to traditional SaaS companies.

Scaling Law

concept that shows how model performance typically improves with increases in model size, dataset size, and computational power used for training.

SmolLM

A family of small and efficient language models developed by Hugging Face.

SOTA (State-of-the-Art)

The highest level of performance currently achievable on a specific task or benchmark within a given field.

Synthetic Data

Data generated by AI models, as opposed to real-world data collected from human activities or natural phenomena.

T

Token

A basic unit of text (often a word or part of a word) used by AI models for processing and generation. The cost of using AI models is often expressed in terms of price per token (or per million tokens, Mtokens).

Transformer (Architecture)

A neural network architecture that has become dominant in NLP and is widely used in other AI applications. It relies on the "attention mechanism" to process data.

U

UI (User Interface)

The means by which a user interacts with a software application or device.

UX (User Experience)

The overall experience a user has when interacting with a product or service, encompassing usability, accessibility, and satisfaction.

VLM (Vision Language Models):

AI systems that can understand and process both visual (image) and textual information.

Optimists for change

Sia is a next-generation, global management consulting group—born digital, augmented by data, enhanced by creativity, and driven by responsibility. We partner with clients to resolve challenges and capitalize on opportunities. We believe that in today's world of change and disruption, optimism is a force multiplier.