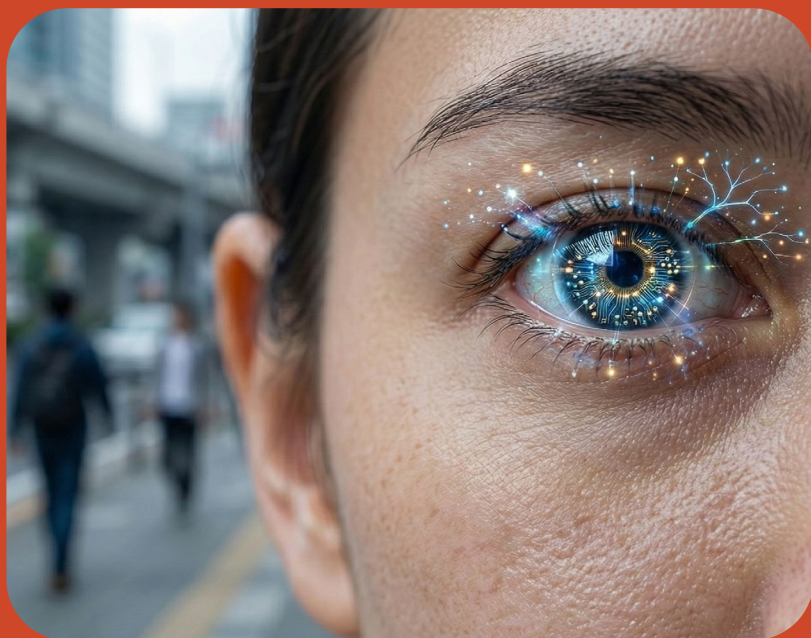


INVESTNL



# AI Deep Dive: Strategic investing in the age of intelligence



In cooperation with:

**ROM ♦ Nederland**

# Contents

Foreword	5
Executive summary	6
Scope	8
<b>1. Fundamentals</b>	<b>9</b>
Machine Learning	
How models learn	
Why has machine learning taken off in recent years?	
Deep Learning	
Graphics Processing Units	
Transformers	
Foundation Models	
<b>2. Data</b>	<b>15</b>
What data is there?	
Peak data	
Multimodal Foundation Models	
The curious case of language	
Experiential AI	
World models	
Reinforcement learning and open-endedness	
Scientific AI	
Ubiquitous AI	
<b>3. Compute</b>	<b>22</b>
Moore revisited	
The AI technology stack	
Energy	
Energy efficiency	
Hardware efficiency	
Software efficiency	
Cross-cutting efficiency	
Edge AI	
Quantum	
Neuromorphic computing	
Efficiency, experience and expressivity	

<b>4. Models</b>	<b>27</b>
Post-training	
Reasoning and the inference scaling paradigm	
Open source	
Continual and online learning	
Small models	
Agentic AI	
Artificial General Intelligence	
Integration	
<b>5. Market</b>	<b>33</b>
AI usage	
Regulation	
Global market	
Private Investments	
European Market	
<b>6. The Netherlands</b>	<b>39</b>
Energy	
Knowledge ecosystem	
Talent	
Startups	
Application Domains	
Funding	
Defensible positions	
<b>7. Opportunities</b>	<b>46</b>
Vision	
Investment opportunities	
Government	
Scenarios	
<b>Afterword</b>	<b>52</b>
<b>Annex I: Investment hypotheses</b>	<b>53</b>
<b>Annex II: Notes on methodology</b>	<b>54</b>
<b>Annex III: Figure references</b>	<b>55</b>
<b>About the author</b>	<b>56</b>

# Foreword

It is with a mix of pride and mild astonishment that we present the fifth edition of the Invest-NL Deep Tech Fund Deep Dive series. Five already. Apparently, time accelerates when you're examining technologies that also accelerate everything else. This edition's topic, Artificial Intelligence, could hardly be more fitting. Few technologies have moved from "interesting" to "inescapable" quite so quickly.

This Deep Dive is the result of an exceptionally close collaboration between Invest-NL and ROM Nederland. And "close" is not just polite phrasing; this publication truly could not have come to life without our joint effort, shared mission, and countless discussions fueled by urgency and curiosity.

A very special word of thanks goes to Stefan Leijnen, the author of this report. Stefan has delivered a masterwork: comprehensive without being abstract, sharp without being heavy, and deeply insightful without losing sight of what really matters for the Dutch AI ecosystem. It is no small feat to turn the entire global AI landscape into something both readable and genuinely strategic. But he did.

I am also deeply grateful to Arjan van den Born of ROM Nederland, whose enormous knowledge - and even larger network - provided foundations and perspectives that strengthened this report in countless ways. And to my colleagues Liz Duijves and Elise Lie, whose tireless dedication, sharp judgment, and relentless drive were essential in turning this Deep Dive into the report you are now reading. Without these people, this document would simply not exist in its current form.

So what is the essence of this Deep Dive? In short: AI is no longer a single technology. It is the new production factor of the 21st century. The Netherlands cannot - and should not - try to beat the U.S. or China in sheer scale. But we can win in the places where scale is not the winning strategy: where deep expertise, trusted infrastructure, scientific strength, and defensible data pipelines create lasting strategic advantage. The report identifies five such opportunity areas where the Netherlands can build real leverage. And these are not just opportunities; they are strategic control points of the emerging AI economy.

Which leads me to an important call to action. If the Netherlands and Europe want to remain economically competitive, technologically sovereign, and aligned with the values we hold dear, now is the moment to act. We invite policymakers, investors, entrepreneurs, research institutes, corporates, and the broader AI ecosystem to join forces. To invest more ambitiously. To collaborate more openly. To think more strategically. And, occasionally, to dare more boldly.

Gert-Jan Vaessen  
Fund Manager, Deep Tech Fund  
Invest-NL

# Executive summary

**Advances in machine learning, foundation models and energy-efficient computing are turning AI into a general-purpose technology with transformative impact across industries, sciences and public sectors. Information work is progressively being automated. The Dutch economy is at the cusp of major changes in the way work is organized. Nations capable of developing, scaling, and governing advanced AI systems increasingly shape global power dynamics; economically, militarily, and normatively. If the Netherlands fails to manage the rapid uptake of AI properly, the consequences could be severe.**

Trying to close the innovation gap with the U.S. or China in areas where they are winning today is a losing strategy. Our capital markets are comparatively shallow, we lack domestic technology giants capable of investing hundreds of billions of euros annually in AI infrastructure and research, and we are late in articulating an industrial policy tailored to the new geopolitical and technological era. AI markets display strong winners-take-most dynamics, as advantages accrue to companies with access to proprietary data and strong distribution channels. In this environment, strategic positioning must guide investment decisions. The Netherlands can secure a strategically differentiated position by doubling down on its structural strengths, anticipating likely technological directions, taking calculated risks, and investing in radical innovation.

AI model quality ultimately depends on access to high-quality and high-volume data, and the compute capacity necessary to process it. The growth of public text datasets is flattening and being overtaken by richer data types such as camera and sensor streams, laboratory and scientific instruments outputs, and robotics-generated data. Though GPUs are likely to remain the dominant paradigm in compute for the near future, alternative AI chips are evolving driven by potential gains in energy-efficiency and local computation.

Where models and infrastructure commodify, competitive advantage will increasingly stem from access to unique, high-quality, trusted data, making it strategically important to control future data pipelines, human and real-world interfaces, and markets for data sharing and model validation. Where scaling laws for models show diminishing returns, opportunities emerge in smaller, high-quality models and edge-deployed systems that leverage Europe's advantages. Integrating models into business contexts, technical architectures and usable interfaces can unlock differentiated value.

This report identifies five high-potential opportunity areas where investors can capture value, build defensible positions, and secure national interests:

- 1 Vertical applications in strategic sectors, that combine global market demand with defensible positions to build proprietary data pipelines and strong adoption potential. In the Netherlands these sectors are agrifood, logistics, energy, high-tech manufacturing and healthcare.
- 2 Human and real-world interfaces, as the growth of data from intelligence devices will outpace the speed at which humans generate text. Leveraging Dutch leadership in responsible AI, companies that develop AI's future interfaces secure strategic positions in data pipelines.

- 3 Data-sharing and model-validation platforms, as proprietary and continuously refreshed data loops become strategic assets. Structural financial incentives aligned with European legislation could motivate businesses to share data and help reduce fragmentation of data accessibility.
- 4 Scientific AI, as discovery in material, physics, geo, and life sciences is shifting toward model-driven, simulation-rich workflows. Significant opportunities are to be found where collaborations can be forged with leading Dutch academic research groups and scientific data sets.
- 5 AI-accelerators, photonics and neuromorphic hardware, as a hardware breakthrough could yield substantial energy-efficiencies and upend the GPU-dominated AI chip market. Building on its strong existing semiconductor ecosystem, the Netherlands could emerge as a global AI leader in energy-efficiency, local computation and analog chip capabilities.

Countries that succeed in aligning their innovation ecosystems, capital markets, regulatory frameworks, and public institutions are best positioned to benefit from AI-driven transformation. As a flanking policy for strategic investments, it is essential that the Dutch government removes structural barriers, including regulatory bottlenecks, high energy prices, and limited space for data centers. To reduce fragmentation across capital markets<sup>1</sup>, usage markets, and data accessibility it is necessary to coordinate at a European scale. The government must act with an AI industrial policy that mobilizes private investments, coordinates public policy areas in a European context, builds and funds hardware-software aligned AI roadmaps to overcome infrastructure and innovation gaps, and makes difficult choices based on strategy and foresight.

This industrial policy should be executed with sufficient public funding and scale specific instruments. Increased access to tickets in the range of €20–30 million is crucial for many Dutch AI startups to cross the gap to profitability. Ticket sizes of €100+ million are needed to grow domestic champions capable of competing globally in hardware and foundation models. This should be reinforced by strong public demand and

willingness to adopt AI. Challenge-based financing allows public and private organizations to fund breakthroughs and become launching customers for solutions.

AI should fundamentally be a public-private technology. The AI industrial policy needs to be backed by a strong public-private investment agenda, creating scale and removing barriers for pension funds, insurers and banks to participate in long-term, high-risk AI investments. As a supportive pillar of its industrial policy, the Netherlands could invest in the AI Gigafactory of the future: a large-scale distributed deployment network optimized for edge-AI, energy-efficient hardware, and privacy-preserving local compute. To protect sensitive data and vital infrastructure for its citizens and businesses, the Netherlands should build public institutions that bring sensitive and vital data and digital infrastructure under democratic governance.

Only through a coordinated, long-horizon approach can the Netherlands build AI capacity that is globally competitive, aligned with democratic values, and capable of supporting long-term prosperity and security.



Only through a **coordinated, long-horizon approach** can the Netherlands build AI capacity that is globally competitive, aligned with democratic values, and capable of supporting long-term prosperity and security.

1. [https://commission.europa.eu/topics/competitiveness/draghi-report\\_en](https://commission.europa.eu/topics/competitiveness/draghi-report_en)

# Scope

**This report highlights technological and economic opportunities in AI and aims to provide a knowledge base to guide investment and policy decisions. It is written for investors, policymakers, and innovators to create a shared understanding of key developments in AI. It intends to stimulate strategic investment and accelerate innovation and scale-up of AI technologies in the Netherlands.**

The report is written against the backdrop of three intersecting concerns regarding a Netherlands' AI industrial policy: economic competitiveness, strategic autonomy, and alignment with European values. While these dimensions reinforce one another, the primary focus for this report is economic: identifying investment opportunities for the Netherlands in the AI value chain over the next three to five years.

Three clarifications frame this scope. First, the report offers a moment-in-time perspective on a field characterized by extremely rapid change. Second, the report offers an integrated look at AI software and hardware. Third, the aim is not to present a conclusive blueprint, but to stimulate informed debate about where the Netherlands should prioritize investment and coordination.

Chapter 1 introduces key AI concepts and developments of recent years; readers familiar with the field may choose to skip it. Chapters 2 through 6 provide a technical and economic assessment of data, compute, models, markets, and the Netherlands' position within these dynamics.

The final chapter presents an integrated analysis and a set of subjective but grounded investment opportunities for private and public investors, and the government of the Netherlands. This chapter can be read standalone, where its underpinnings can be found in the preceding chapters.

Ultimately, the aim of this report is to offer clarity, provoke discussion, and support strategic decision-making at a moment when AI's trajectory, and our ability to shape it, is far from predetermined.

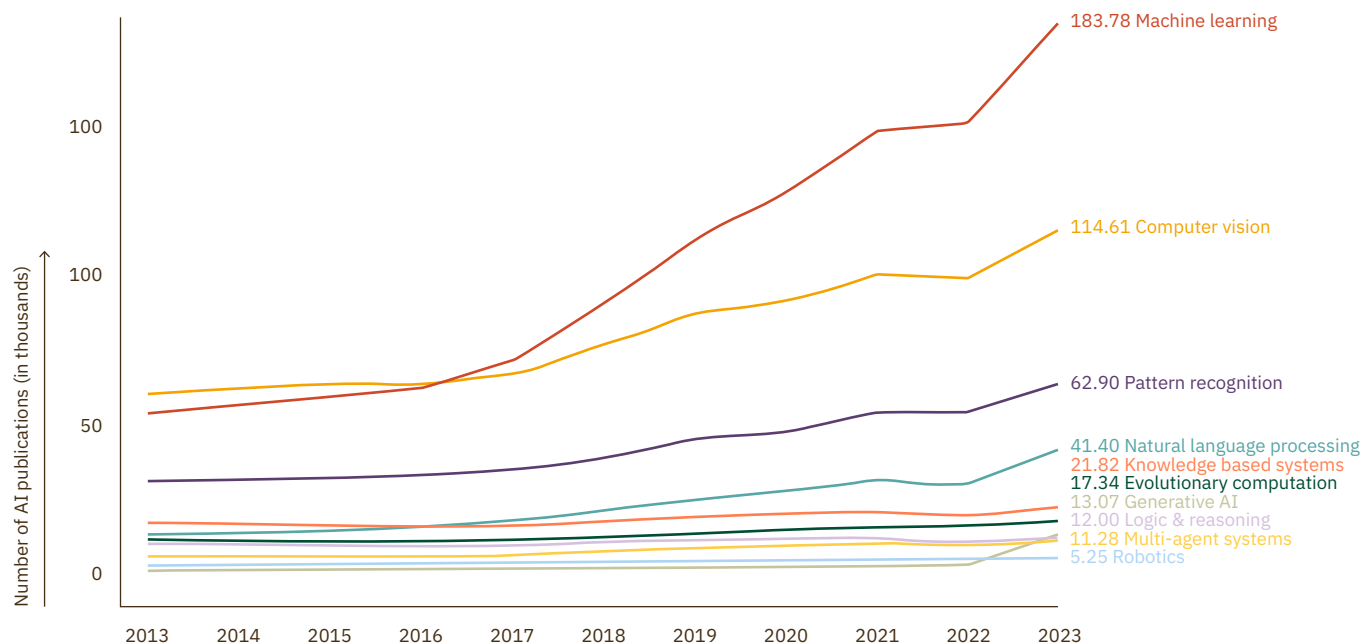


The scope of this report is to **offer clarity, provoke discussion, and support strategic decision-making** at a moment when AI's trajectory, and our ability to shape it, is far from predetermined.

# 1. Fundamentals

**Artificial Intelligence is the ambition to make machines that are intelligent, or appear to be. This ambition is older than the machines that we see around us today<sup>2</sup>. Our intelligence is what sets us humans apart, and people have dreamt for ages of building machines that can do our work for us, are creative like us, that help us understand ourselves, and that make the universe a less lonely place.**

Over the centuries, technological innovations such as hydraulic systems, clockworks and steam engines have each been considered as artificial intelligence. More recently, breakthroughs in information processing, connectivity and algorithms have fueled this ambition, redefining the field of AI. Future breakthroughs will yet again redefine the field. Today, progress in AI is primarily driven by advances in machine learning.



**Figure 1. Number of AI publications by select top topics, 2013-23. Source: AI Index Report, 2025**

<sup>2</sup> <https://www.wired.com/story/opinion-ai-is-an-ideology-not-a-technology/>

# Machine Learning

Machine learning enables computers to identify patterns and make predictions, without explicit programming. A computational model is refined over time by exposing it to data; when there is a sufficient fit to reality, the model can make accurate predictions about new, unseen data. These predictions can be about anything the data at hand allows: the next day's weather, what an image of a cat looks like, what an answer to a question should be, where a robot should move. Where computer programming can be understood as accelerated logic, so can machine learning be understood as accelerated statistics. The strength of machine learning lies in its power to scale, and on its general applicability: many, if not all, patterns found in nature may be efficiently discovered and modeled by learning algorithms<sup>3</sup>.

In theory, using machine learning instead of programming means that the complexity of what you can create is no longer bound by your own intelligence - as you no longer need to understand it to build it - and you can create models of infinite complexity and predictive accuracy. In practice, limitations do exist: data availability, time, processing power and energy. Also, because you no longer need to understand how the model works, there are practical limitations in explaining and controlling the outcomes of machine learning. Many efforts in the field of AI in the age of machine learning are about dealing with these limitations.

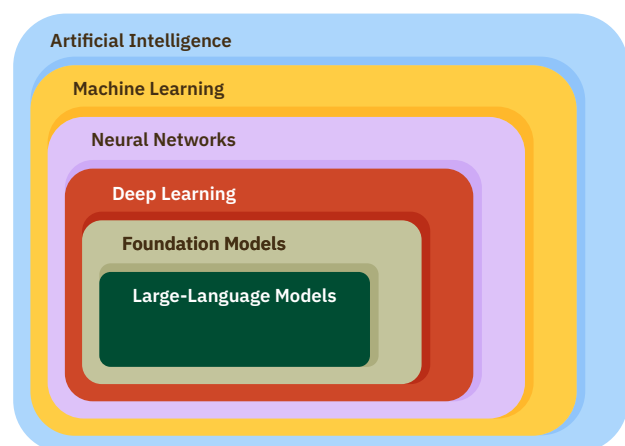


Figure 2. AI technology terms

# How models learn

In machine learning, how a task is performed is inferred from examples<sup>4</sup>. Learning approaches are generally divided into four categories: supervised, unsupervised, semi-supervised, and reinforcement learning. Supervised learning relies on large volumes of labeled data to train models that can classify, predict, or estimate outcomes—for example, detecting fraud in financial transactions or forecasting demand in supply chains. In contrast, unsupervised learning works with unlabeled data, identifying hidden patterns or relationships without predefined outcomes, for example in market segmentation, anomaly detection, and data compression. This approach is valuable in scenarios where labeled data is scarce.

Semi-supervised learning leverages a small set of labeled examples alongside a much larger pool of unlabeled data, a practical approach where labeling is costly or limited. Finally, reinforcement learning focuses on decision-making through trial and error, with algorithms learning strategies by maximizing long-term rewards. This method underpins breakthroughs in robotics, autonomous systems, and is increasingly applied where data becomes sparse.

What the algorithm is trying to optimize is defined by an objective function (also called loss function or reward function). For supervised learning, it measures the error between the model's predictions and the true labeled outputs. For unsupervised learning, the objective function captures how well the model discovers structure in unlabeled data. Since there are no labels, the function measures internal criteria like compactness, separation, or reconstruction accuracy. In semi-supervised learning, it balances supervised error from labeled data and unsupervised error from unlabeled data. In reinforcement learning the objective function maximizes the expected cumulative reward over time, to learn a strategy that maximizes long-term outcomes based on interaction with an environment.

For large models, such as large-language models, a combination of semi-supervised learning and reinforcement learning is often used, due to the vast amounts of unlabeled data available for training them and the opportunities for interacting with users.

3. <https://www.nobelprize.org/uploads/2024/12/hassabis-lecture.pdf>

4. <https://fsi.stanford.edu/publication/opportunities-and-risks-foundation-models>

# Why has machine learning taken off in recent years?

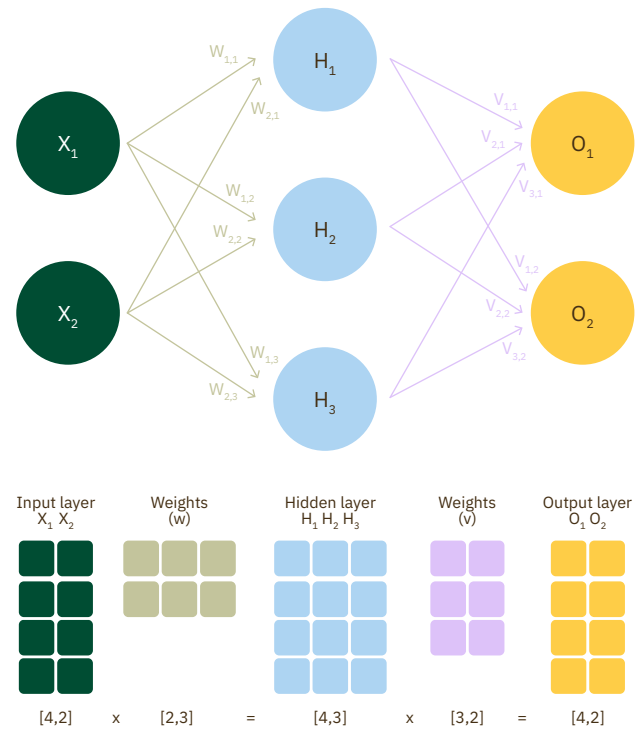
Some of the core concepts that make up today's machine learning systems have been around for decades; the technology needed more computation and data than was available at the time. Neural networks, the model architecture of choice for many machine learning systems today including large-language models, were invented almost eighty years ago. In recent years, machine learning has accelerated dramatically due to the convergence of several key breakthroughs. The exponential growth of the internet has created vast datasets that fuel learning algorithms, making large-scale and semi-supervised learning feasible, initially enabled by deep learning and GPUs, described below. The introduction of transformer architectures further amplified this trajectory by enabling models to capture long-range dependencies and context in data, leading to breakthroughs in language models and showing that machine learning is not confined to narrow tasks but can adapt across domains.

## Deep Learning

Neural networks are computational models inspired by how biological neurons process and transmit signals. They consist of interconnected layers of nodes (artificial neurons), where each nodes performs a simple mathematical operation, such as taking input values, multiplying them by adjustable weights, adding a bias term, and passing the result through a non-linear activation function. Training a neural network involves adjusting the weights and biases so that the network's predictions match desired outputs, a process typically guided by gradient-based optimization using supervised, unsupervised, semi-supervised or reinforcement learning. What the desired outputs are is defined by the objective function.

At its core lies the insight that even a simple neural network with a single hidden layer is a universal function approximator: a model capable, in theory, of learning any regular, predictable relationship between inputs and outputs, no matter how complex, given enough neurons. However, these shallow networks are inefficient and computationally intractable for real-world problems. They would require an impractically large number of neurons and parameters to capture high-dimensional relationships such as those in

images or natural language. This bottleneck has been solved by stacking many hidden layers (hence "deep" learning) in a neural network, which allows much more compact representations.



**Figure 3. A neural network with input (X) nodes, hidden (H) nodes and output (O) nodes can be modelled as a series of matrix multiplications** Source: Machine Learning, Lecture 6: Deep Learning, Anshumali Shrivastava, 2022

## Graphics Processing Units

The computations needed for deep learning can be expressed elegantly as matrices and tensors (multi-dimensional arrays). They can be efficiently modeled as multiplications of input vectors with a weight matrix and adding a bias vector, followed by a non-linear activation function. Stacking these operations yields matrix chains that map input data to predictions. For images, video, or other multidimensional data, the same principle applies, enabling efficient representation of complex structures like color channels, time sequences, or 3D volumes.

This tensor-based formulation is key to modern deep learning: it not only makes the mathematics compact and scalable but also aligns with graphics processor unit (GPU) hardware. Originally developed for high-performance computer

graphics, these processors allow for massively parallel arithmetic operations on tensors<sup>5</sup>. The availability of these operations made GPUs readily available for deep learning at significantly better performance compared to CPUs. Whether the math behind graphics and deep learning happens to be similar by accident or whether this similarity unveils something about the structure of reality is a question relevant for quantum, photonic and neuromorphic hardware (see chapter 3. Compute).

This scalability has enabled the rise of huge pre-trained models like BERT and GPT, which learn transferable knowledge from massive datasets. Their flexibility has led to dominance not only in natural language processing but also in vision, multimodal, and scientific domains.

## Transformers

Much of AI research and industry deployment has standardized around transformers.

Over the past decade, model design in deep learning has moved from a diverse set of domain-specific architectures toward a more homogenized landscape dominated by a few highly effective approaches. Early breakthroughs in deep learning were domain-specific, e.g. convolutional neural networks (CNNs) for vision, recurrent neural networks (RNNs) and LSTMs for sequence modeling.

Over the past decade, model design in deep learning has moved from a diverse set of domain-specific architectures toward a more homogenized landscape dominated by a few highly effective approaches (see figure 4 on the next page).

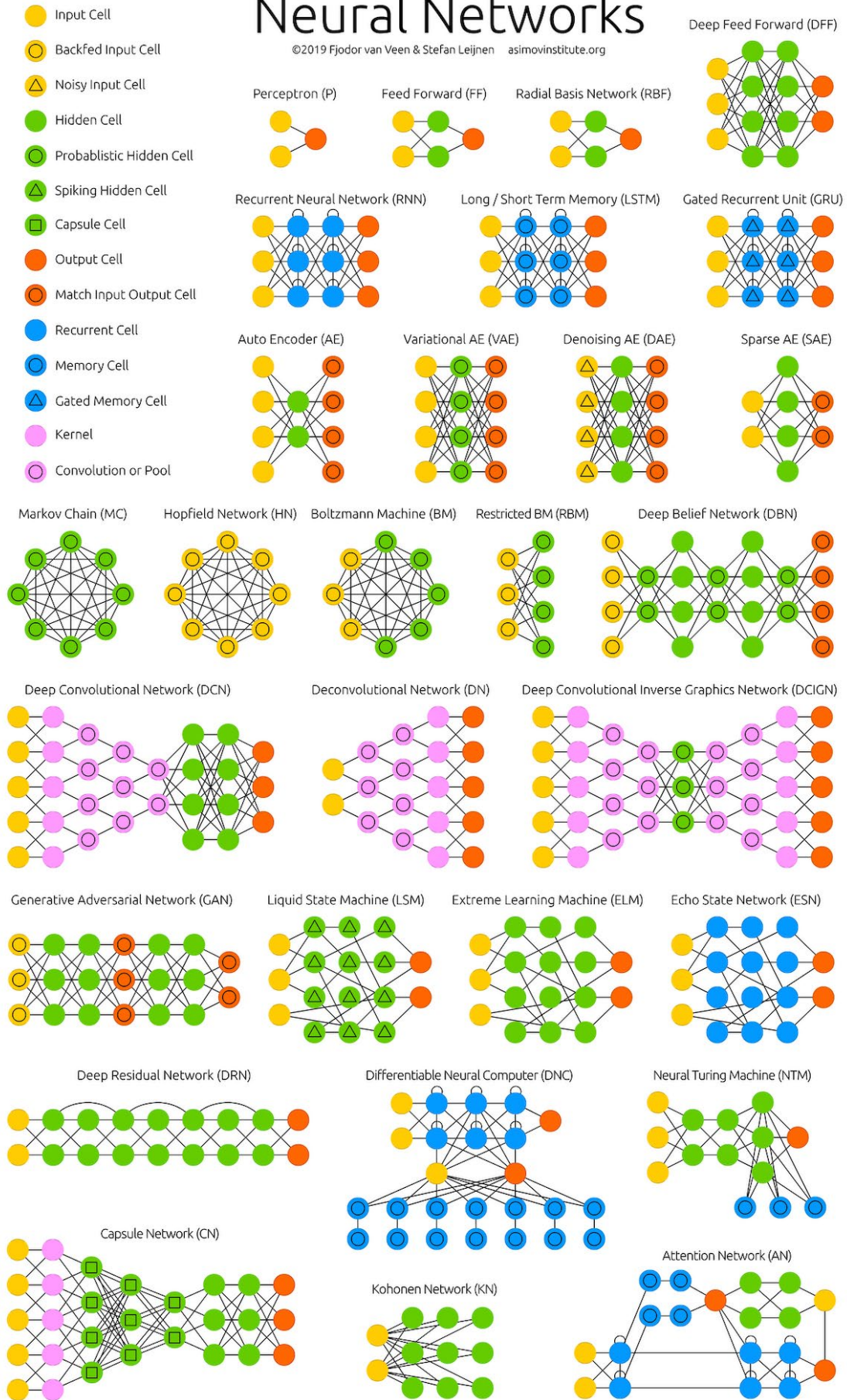
The introduction of the transformer architecture in 2017 reshaped the model landscape. Unlike RNNs, which handle tokens step by step and therefore train slowly, transformers leverage self-attention to process all tokens simultaneously, enabling massive parallelization and faster training on modern hardware. Multi-head attention further enriches representations by letting the model focus on multiple aspects of context at once. They also capture long-range dependencies better: where RNNs struggle with vanishing gradients and compressed hidden states, transformers learn relationships between all elements of a sequence simultaneously, eliminating the bottlenecks of recurrence and enabling efficient training on GPUs.

---

5. e.g. matrix multiplication, matrix convolution, intra-warp communication for reductions, softmax and attention mechanisms

# A mostly complete chart of Neural Networks

©2019 Fjodor van Veen & Stefan Leijnen asimovinstitute.org



**Figure 4. The neural network zoo showing the evolution of deep learning architectures, with the Transformer (Attention Network) architecture at the bottom-right. Source: The Asimov Institute**

# Foundation Models

Foundation models are very large deep learning systems trained on vast datasets, enabling them to be adapted across a wide range of tasks and modalities. They are the driver of generative AI, capable of producing highly complex output. Frontier foundation models like OpenAI GPT-5 and Google Gemini are estimated to consist of hundreds of millions to several billions of artificial neurons and over a trillion connections, also called parameters. For comparison, the human brain has ~86 billion biological neurons, though these are more complex and interconnected than artificial neurons.

The scale of foundation models give rise to advanced emergent functionalities that materialize as these models scale, enabling capabilities such as translation, summarization, code generation, and

image editing. Given the significant time and cost required to train foundation models, it is infeasible to retrain and test them from scratch for every new use case. Instead, a general-purpose foundation model can be adapted for specific applications while retaining its broad capabilities. Post-training methods such as finetuning allow these pre-trained models to be adapted, so they behave helpfully, safely, and appropriately for target domains. Due to the relative high cost of training them, foundation models are valuable assets. As these models sit at the core of training, post-training, and inference pipelines, they have become a new component of the AI supply chain.

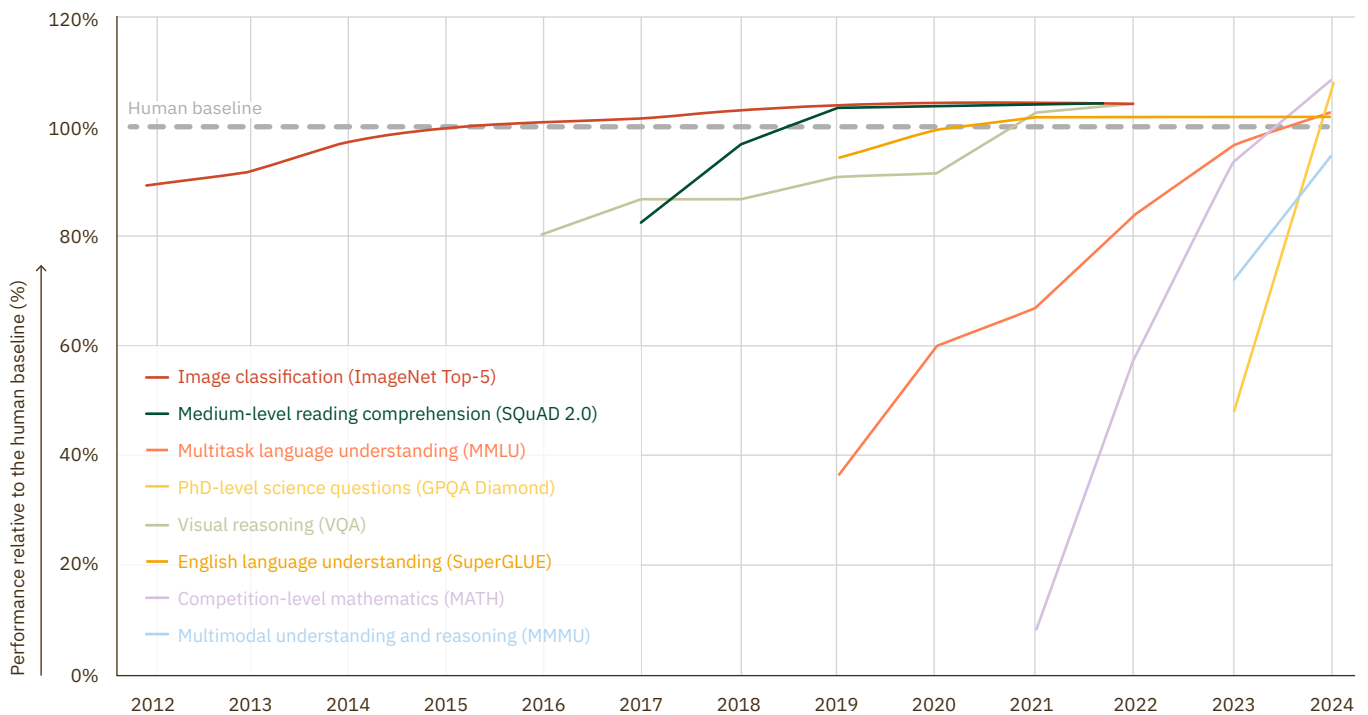


Figure 5. Select AI Index technical performance benchmarks vs human performance. Source: Epoch AI, AI Index Report, 2025

## 2. Data

Creating a machine learning model requires an engineered pipeline, typically spanning data ingestion; data validation to catch anomalies; data preprocessing and feature engineering to centralize and reuse features; model training; post-training steps such as fine-tuning or distillation for target domains or safety alignment; deployment for model inference; and monitoring for latency, error rate or model drift.

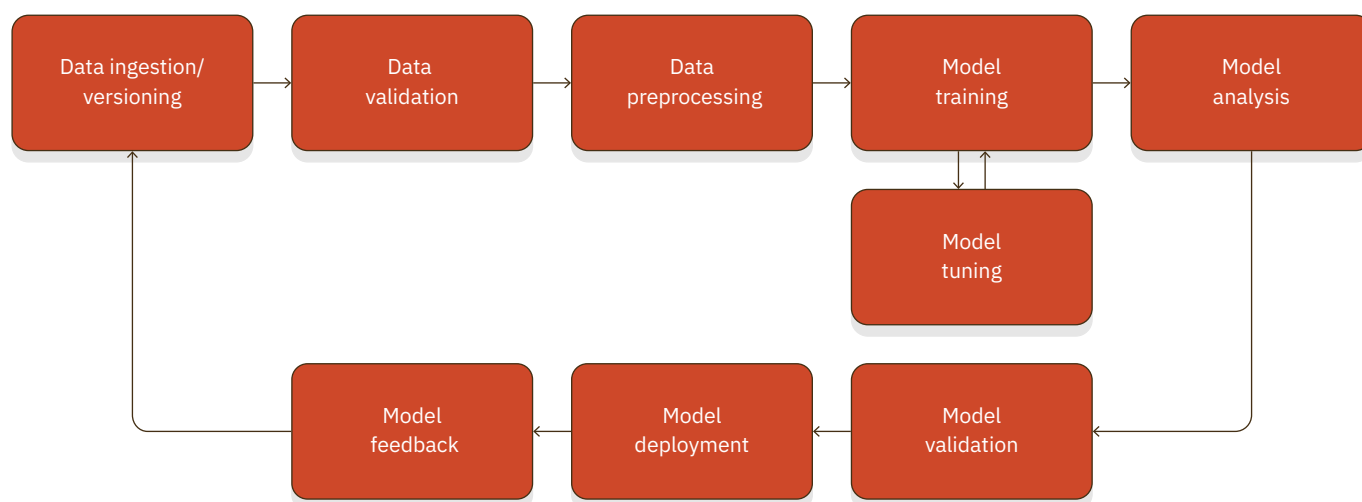


Figure 6. Steps to building a model

Training, post-training, and inference usually draw the most attention, since they are heavily constrained by computing resources, drive hardware and software innovation, and determine how quickly a model converges toward optimal predictive capacity. However, it is ultimately the input to this pipeline, the underlying data, that sets the ceiling for the model's performance. The potential quality of a model is determined by data; how fast this potential is approached is determined by the available compute, in terms of energy, hardware and software.

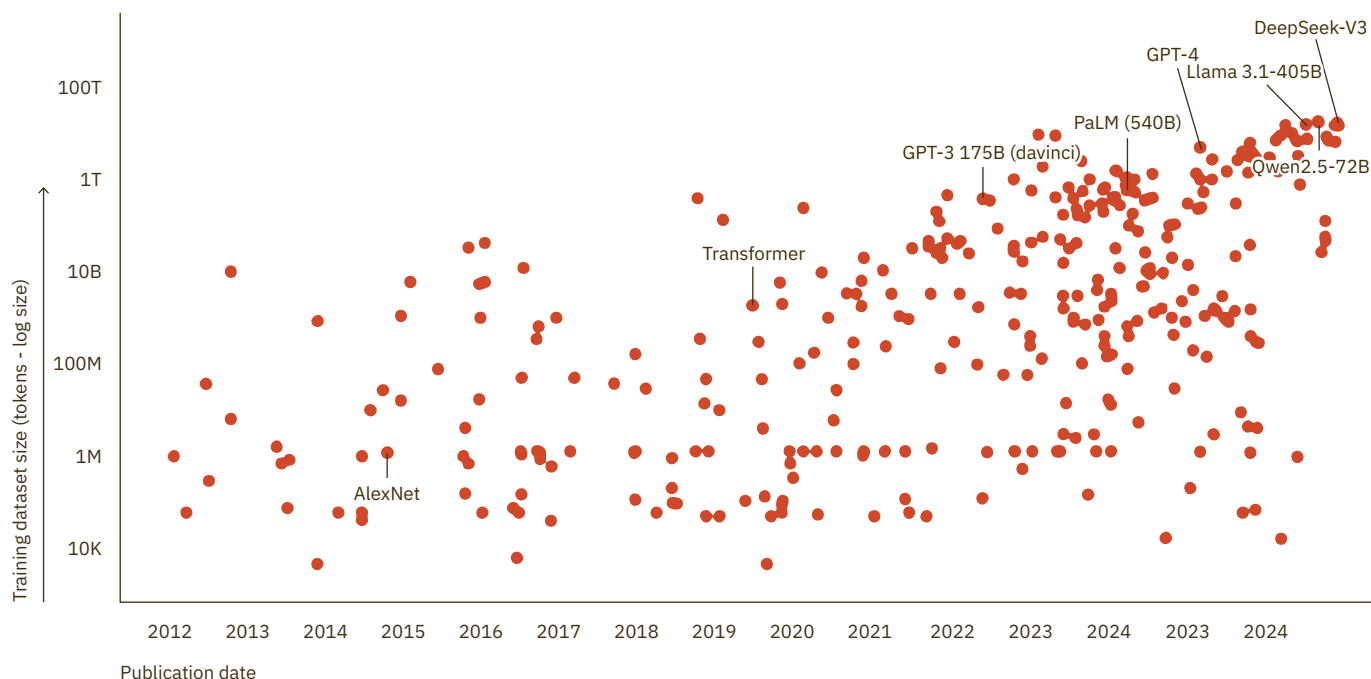
Models can be understood as lossy compressors of the data they ingest. They distill statistical structure in data into the model parameters and will only

generalize along patterns present or implied in that data. What a model can do therefore mirrors what data we choose to measure and store.

### What data is there?

This year alone, the total amount of data generated globally is estimated at 175 zettabytes (175x10<sup>21</sup> bytes, or 43750 billion DVDs), up from 44 ZB in 2020, and 2 ZB in 2010. This data consists of text (web pages, documents, chats, code), images and videos (consumer, satellite, surveillance) accounting for 80% of the total data volume, audio (speech, music), tabular data (transactions,

logs), geodata (GPS, weather, traffic), behavioral data (clickstreams, purchase histories, social network graphs), health data (electronic health records, wearables), industrial data (aerospace and manufacturing sensor data), scientific data (lab instrumentation), and synthetic data (simulations, generative AI).



**Figure 7. Training dataset size of notable AI models, 2010-24.** Source: Epoch AI, AI Index Report, 2025

Data growth is increasingly driven by streaming and machine-generated data sources. Internet of Things (IoT) endpoints alone will generate data in the order of 80 ZB this year. Because human-generated data is bound by our physical and mental capacity to interact and make decisions, data generated by sensors and devices are expected to outgrow human-generated data by orders of magnitude.

## Peak data

Though the stock of human-generated public text is growing, it is outpaced by the growth of dataset sizes used for training LLMs. The term Peak data signifies the era of endless gains from ever-larger, internet-scraped training datasets is ending, as the quantity and quality of human-generated data available for AI models is reaching saturation<sup>6</sup>. After

using the relatively open dataset made available by the internet, we may have exhausted the fossil fuel of AI, and future improvements will struggle to find novel, high-signal data. Relying on synthetic data as a substitute could lead to model collapse, where errors and biases amplify over generations of models trained on machine-generated content.

Peak data may frame a turning point in AI development: from ever-expanding data scale to more careful curation, efficiency, synthetic augmentation, and new paradigms for learning. It also underscores the importance of proprietary data and data-generating pipelines. In a world where every competitor has access to the same foundation models vertical, domain-tuned, and continuously refreshed feedback loops are a defensible moat and a growth engine. Data ownership gives organizations a sustainable competitive edge, since competitors cannot easily replicate the unique insights derived

<sup>6</sup> <https://www.theverge.com/2024/12/13/24320811/what-ilya-sutskever-sees-openai-model-data-training>

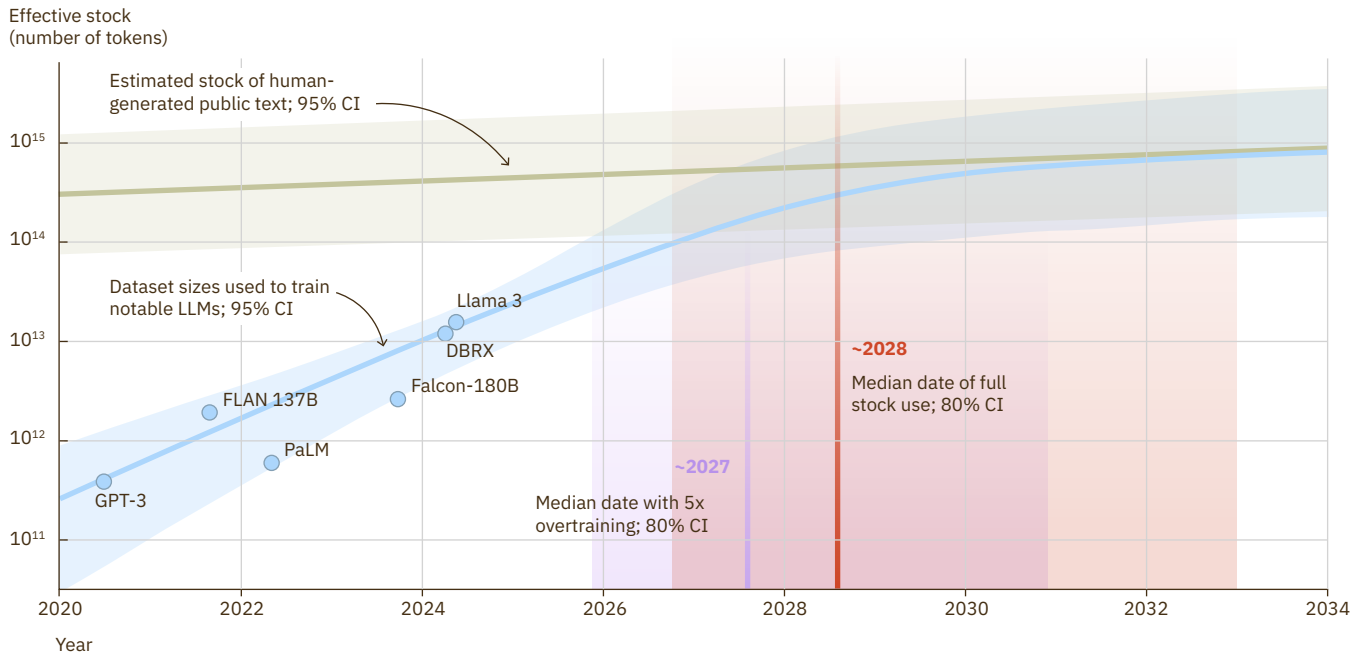


Figure 8. Projections of the stock of public text and data usage. Source: Epoch AI, 2025

from them. It also reduces dependence on publicly available or third-party datasets, which may be scarce, expensive, or subject to shifting access rules.

## Multimodal Foundation Models

Growth of human-generated text is not only outpaced by growth of dataset sizes used for training LLMs, but also by the growth of data in other modalities. Unlike LLMs, which only process and generate text, Multimodal Large Language Models (MLLM) can handle multiple input and output modalities such as text, images, audio or video, while Omni-modal models can handle any combination of data modalities. This allows them to not only read and write language but also interpret visual content, describe scenes, answer questions about images, transcribe or generate speech, and integrate information across modalities. MLLMs and Omni-modal models expand the versatility of LLMs, for example, enabling AI to analyze a chart, explain it in natural language, and then answer follow-up questions about it. They are foundational for tasks that require grounding language in the sensory world, such as vision–language reasoning, medical imaging, and multimodal search. In robotics, Vision-Language-Action Models (VLAMs)

extend this further by not only interpreting and generating information from images and text but also producing actions that can be executed in an environment. For example, a VLAM might take a camera feed (vision), parse a human instruction (language), and then output a sequence of motor commands (action) to control a robot arm or navigate a digital environment. Though there is a dependence of these models on LLMs and therefore language data, due to the difference in data growth these models can be expected to increasingly rely on non-language data modalities.

## The curious case of language

From the early days of computing, language has played a central role in the development of artificial intelligence. The famous Turing Test<sup>7</sup> claims that the ability of a machine to hold a convincing conversation should be seen as evidence of intelligence. In 2022, the so-called LaMDA incident rekindled debate about whether machines could truly understand language, when a software engineer at Google claimed their LaMDA language model had become sentient<sup>8</sup>. When ChatGPT was launched in November 2022, it was framed as a chatbot. However, language is not

7. <https://plato.stanford.edu/entries/turing-test/>

8. <https://www.wired.com/story/lamda-sentient-ai-bias-google-blake-lemoine/>

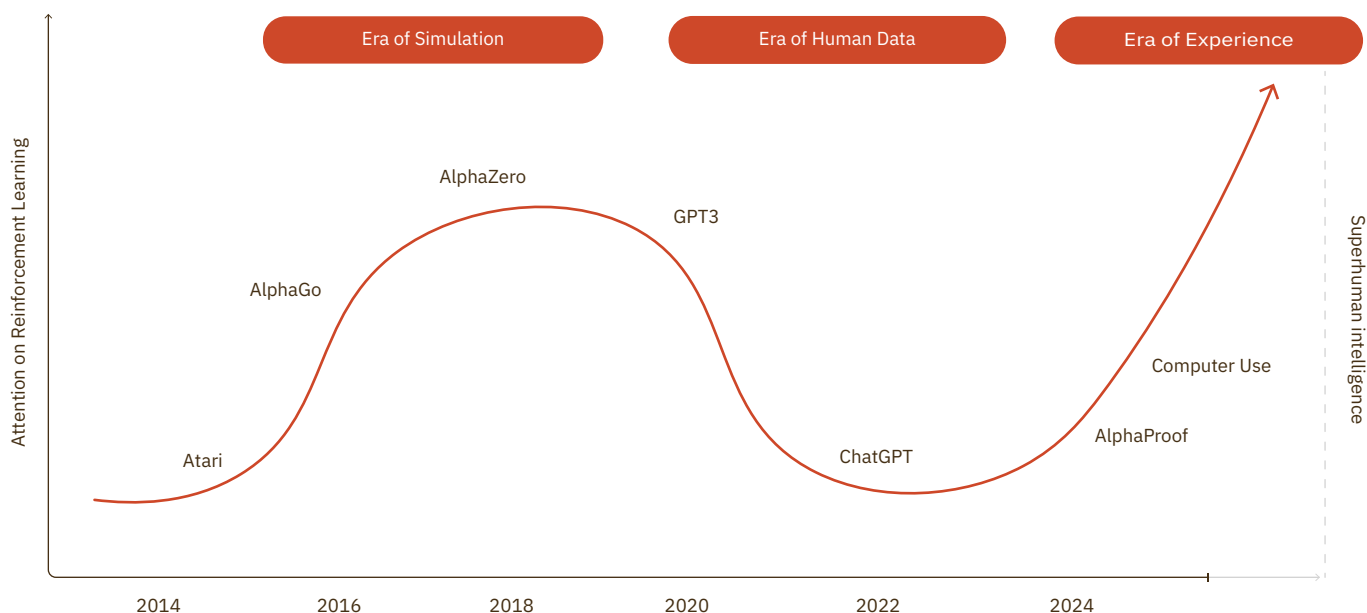
just a medium of communication but a tool for reflection, reasoning and abstraction. Whether these capabilities are intrinsically tied to language or if can equally emerge from other data modalities, is an open research question. It is not yet clear whether language is truly unique in this respect. Yet, the hypothesis that language is a special case has already attracted extraordinary amounts of investment capital, even given the limitation imposed by the slow pace of language data growth. At the current pace of progress, we may learn whether this hypothesis is true or false before we fully understand why.

example is DeepMind’s AlphaGo, the AI that defeated world champion Lee Sedol in the game of Go<sup>10</sup>. In interaction, data is produced by actions and observations under controlled settings where consequences can be simulated, allowing the model to explore novel state-action trajectories. Because this data is generated, it can in principle be tailored, replayed, or amplified in ways that real-world data cannot, and can adapt as the model improves.

## Experiential AI

Peak data points to transition from models primarily learning from human-generated data to models learning predominantly from experience: models interacting with environments over long, continuous streams, which generates data that scales with the model’s growing capability<sup>9</sup>.

Before the popular emergence of LLM, many foundation models would learn from actively interacting with a simulated environment. A famous



**Figure 9. Rise of Experiential AI.** Source: *Welcome to the Era of Experience*, David Silver & Richard Sutton, 2025

9. <https://storage.googleapis.com/deepmind-media/Era-of-Experience%20/The%20Era%20of%20Experience%20Paper.pdf>  
 10. <https://deepmind.google/research/alphago/>

## World models

In a future era of experience, AI's act in richly grounded environments, not only through human dialogue, and receive rewards derived from real-world signals rather than human judgments, enabling discovery beyond human priors. Planning, reasoning, and internal representations may evolve beyond human modes (e.g. non-linguistic reasoning) to world models: internal representations that help AI's simulate and predict how an environment works and plan actions by imagining future outcomes instead of relying on trial and error<sup>11</sup>. World models model the physical world and logic of action itself, allowing for richer, multi-scale interactions with an environment that is not limited by human preconceptions of what these interactions should be. This shift toward experiential learning promises breakthroughs in autonomy, adaptation, and scientific discovery, while also introducing novel safety, interpretability, and control challenges.

## Reinforcement learning and open-endedness

The era of experience holds promise for reinforcement learning (RL) because it envisions a future where AI is no longer bottlenecked by limited human-curated datasets but can generate virtually unlimited training signals through interaction with their environments. This aligns directly with RL's core strength: learning from trial and error, exploration, and feedback loops. By grounding AIs in dynamic environments with diverse and information-rich reward signals, RL can scale more naturally, enabling the emergence of sophisticated behaviors and strategies beyond what supervised or semi-supervised learning on static data allows.

Open-ended AI systems continually propose and solve new tasks without a fixed endpoint and accumulate the resulting skills. Interactive and persistent world models make this increasingly feasible. Learning continuously from their own experience, RL methods may support adaptation,

long-horizon planning, and the discovery of novel solutions that humans may not anticipate. This this to work, verifiers and reward models are becoming essential for AI development. This is reflected by a growing interest in RL startups. For example, there are 21 Y-combinator startups active in reinforcement learning in 2025, up from 10 in 2024 and 4 in 2023<sup>12</sup>.

## Scientific AI

Scientific research is a major engine of data generation: fields such as genomics, climate science, particle physics, and space exploration rely on massive datasets for simulations, experiments and analysis. Instruments like telescopes, genome sequencers, and particle colliders continuously produce large quantities of data. AI is a powerful tool for science by uncovering patterns data from physics, chemistry and biology that were previously hidden. Companies such as Cradle<sup>13</sup> and Cusp<sup>14</sup> represent a paradigm shift in how research is conducted, combining high-performance computing, closed-loop automated labs, and advanced AI models to accelerate discovery. By enabling large-scale, reproducible experiments and analyzing massive datasets, AI can uncover new patterns and solutions beyond human intuition. Strategic investments in automated experimental laboratories, computing infrastructure, data resources and talent have the potential to democratize access, lower costs, and seed entirely new industries in areas like medicine, materials, engineering and food production<sup>15</sup>. These efforts may leverage known data modalities or introduce new ones like protein or material structures. Foundation models trained on scientific data may be complemented by (M)LLMs that capture patterns in hands-on experimentation and unrecorded tacit knowledge from labs generating unstructured data<sup>16</sup>.

---

11. <https://www.turingpost.com/p/topic-35-what-are-world-models>

12. <https://www.ycombinator.com/companies>

13. <https://www.cradle.bio/>

14. <https://www.cusp.ai/>

15. <https://www.elsevier.com/connect/ai-for-science-a-paradigm-shift-for-scientific-discovery-and-translation>

16. <https://ifp.org/teaching-ai-how-science-actually-works/>

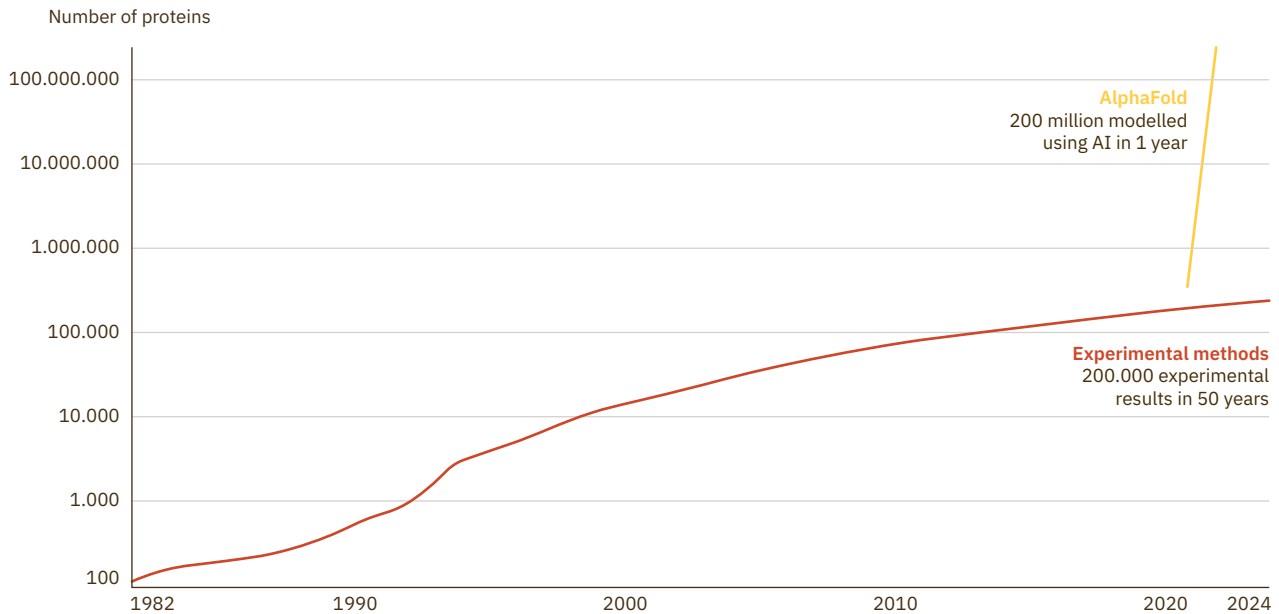


Figure 10. AI modelling accelerated the pace of protein structure analysis by around 45000 times. Source: IEA, 2025

## Ubiquitous AI

There currently are about 50 billion connected devices globally, with the largest categories in industrial, automotive and navigation followed by smart home devices and wearables<sup>17</sup>. IoT devices are expected to generate over 90 zettabytes of data in 2025, out of a total of 175 ZB<sup>18</sup>. Cloud storage currently accounts for most global data, with roughly 60% of corporate data stored in the cloud. Meanwhile, IoT devices and edge computing reduce latency and bandwidth, improving real-time data processing capabilities. Nearly 30% of the world’s data will need real-time processing as computing at the edge continues to grow.

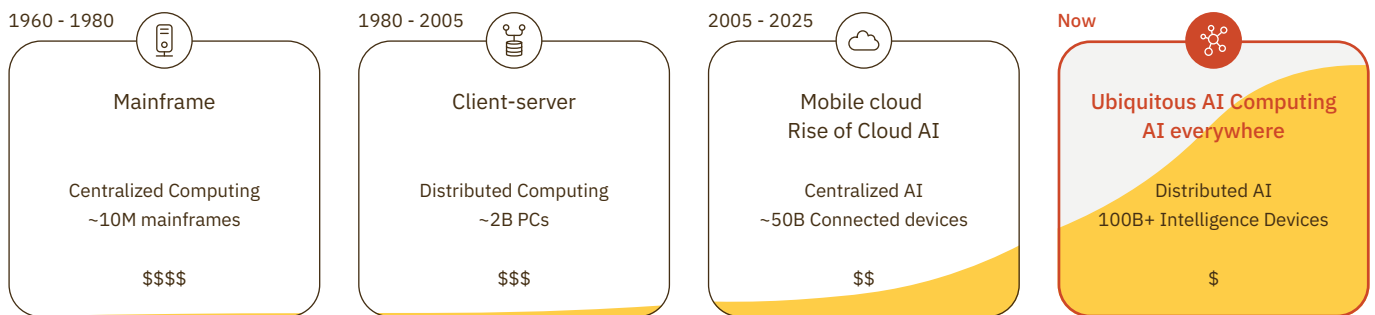


Figure 11. We are moving toward ubiquitous and distributed AI. The growth in data and devices is driven by decrease in AI compute costs, limited scalability of cloud AI, and increase in value created with AI. Courtesy of Axelera AI.

17. IDC

18. IDC & Seagate Data Age 2025, <https://www.seagate.com/files/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf>

As more devices become connected and intelligent and robots are learning to perform novel tasks in never-before-seen environments, AI computing becomes ubiquitous. This trend, combined with experiential AI, holds a powerful promise. These ubiquitous intelligent devices are capable of generating human-generated and non-human generated data at unprecedented volumes. There is strategic value for whoever owns these devices and the gateway to personalized, situated and ubiquitous data gathering. This warrants strategic investments in sensor-carrying interfaces such as autonomous vehicles, wearables, collaborative robots, biometric clothing, smart city infrastructure, and home devices.

Ubiquitous, situated and personalized AI is a candidate interface that may grow to become the primary mode of interaction with AI. Today's dominant interface - typing chat-like prompts in a text box - is a relic of our previous system technology: the search bar is the interface to which we have grown accustomed to for navigating the internet. We can expect more natural and powerful interfaces to AI to emerge and take its place.

# 3. Compute

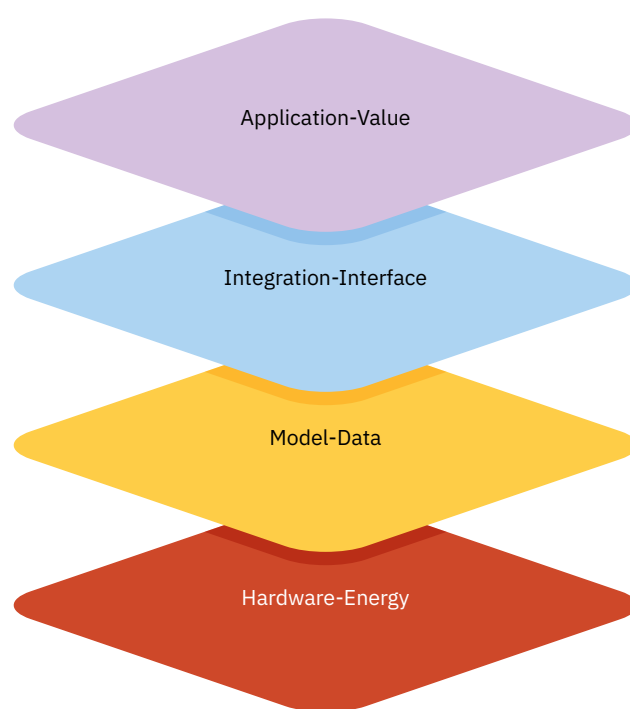
**A central claim of this report is that access to data determines the potential quality of model, while access to compute, in terms of energy, hardware and software, determines how fast this potential is reached. While the global volume of data grows, so does the energy-efficiency of compute. Taken together these two trends yield increasingly powerful models, unlocking improved model performance new capabilities.**

## Moore revisited

Moore’s law, the observation that the number of transistors on a chip doubles roughly every two years while costs fall<sup>19</sup>, has driven progress in computing for decades by delivering exponential increases in processing power and efficiency. Compute is a bottleneck for AI model scaling, and training and inferencing large models require high levels of parallel processing, memory bandwidth, and energy efficiency. Foundation model scaling has made Moore’s law more relevant than ever by amplifying the demand for computational resources and driving innovation for energy-efficient GPUs, TPUs, and other types of AI-specialized hardware. Moore’s prediction about transistor density is evolving into a broad principle about accelerating computational capability, underpinning AI’s rapid advances and shaping the future of hardware development.

## The AI technology stack

AI compute can be represented as a layered technology stack, in order to map investment, regulation, and innovation across the ecosystem<sup>20</sup>. Here we use a four-layered technology stack:



**Figure 12. AI technology stack**

Layers are mutually dependent: applications rely on integration for usability but also shape it; models require infrastructure for training and inference but also define the tensor-operations towards which infrastructure should optimize. It should be noted that this model is a simplification, as each layer could have sub-layers such as orchestration and software optimization. The reason for using these four layers are their distinct possibilities to create defensible positions and their supply chain dynamics.

<sup>19.</sup> and energy-efficiency increases

<sup>20.</sup> c.f. <https://eurostack.eu/>, <https://a16z.com/who-owns-the-generative-ai-platform/>

**The Hardware-Energy layer** captures specialized hardware (CPUs, GPUs, TPUs, networking, storage) that provide the computational power and scalability required for AI workloads. Key companies that operate in this layer include Nvidia, AMD, TSMC, ASML, Google, Intel, and ARM. This layer's key performance characteristic is functional efficiency in terms of energy and time, and involves a complex and often global supply chain, backed by IP-protections such as patents.

**The Model-Data layer** is where models are trained and deployed using vast amounts of data. This layer is the core engine of AI capability, translating raw computational power into statistical representations of knowledge and reasoning. This layer's key performance characteristic is scale: scale of data, computation resources and usage. Scale forms a defensible moat for companies, as holding on to talent and unique knowledge is often transient. Ownership of data and data pipelines also provides a defensible moat. Model ownership tends to be undermined over time by model amortization due to commodification and open-source advancements.

**The Integration-Interface layer** provides tooling, APIs, governance, interfaces, and workflow management needed to operationalize AI systems. Integration enables fine-tuning, security, compliance, and interoperability with enterprise systems. Its key performance characteristics are operational capability and maintainability. This layer offers strategic opportunities for owning data-generating interfaces, such as API's, chatbots, sensors, robots, and integrations with legacy data and systems. Large platforms tend to build entire ecosystem offerings as usage moats while smaller players will find niche opportunities.

**The Application-Value layer** is where user-facing products and services, such as copilots, domain-specific assistants, agents and autonomous systems, leverage the full stack to deliver value to end-users. Its key performance characteristics are user value and market fit. Though the launch of ChatGPT marked a period of general-purpose applications, many future applications are expected to be tailor to a particular niche, company or individual. Applications with vertical integration typically have a defensible moat, i.e. down to the Model-Data layer due to proprietary data and

(finetuned) models, or down to the Integration-Interface layer due to proprietary data-gathering interfaces or devices.

Vertical integration - ownership of aligned products in two or more layers - and horizontal integration - ownership of dominating products within a layer - can yield efficient synergies, for example Google using proprietary TPUs to train its models or Nvidia dominating the GPU market through its CUDA platform. Integration builds strong defensible positions akin to monopolies, which primarily benefits AI incumbents. Leading companies such as Microsoft, Google, and Amazon are strong in end-to-end integration: they own the infrastructure, the platforms, the models, and often the applications built on top. This structure reinforces lock-in, making it difficult for smaller players to compete or innovate independently. Open standards, open-source data and models, interoperable interfaces, and transparent governance limit vertical and horizontal integration and encourage fair competition, reduce switching costs, and foster a healthier market ecosystem.

## Energy

Demand for data center electricity over the next five years will increase due to AI. The International Energy Agency (IEA) has collected recent estimates of expected demand. Estimates for global demand in 2030 range from 200 TWh to over 800 TWh<sup>21</sup>, up by a factor 2x-10x from 2025. Data center capacity of US hyperscalers<sup>22</sup> such as Meta (~9780 MWh), Google (~8960 MWh), Amazon (~7660 MWh) and Microsoft (~6970 MWh) rivals traditional heavy industry in electricity consumption. It is projected that electricity shortages could occur within the next 1-3 years in several major US regions due to unreliability and new AI demand. In comparison, Chinese data center capacity is estimated to be significantly lower with Tencent (~1760 MWh), Alibaba Cloud (~1660 MWh) and Huawei (~1260 MWh), through these figures may be unreliable. Total global data center consumption accounts for ~1.5% of electricity consumption in 2024. The US roughly accounts for 45% of global data center energy consumption, followed by China with 25% and Europe with 15%.

21. <https://www.iea.org/reports/energy-and-ai/>

22. IEA, OMDIA (2025)

# Energy efficiency

Given the growth of energy and compute consumption to match the expected needs of AI, there is a strong incentive for improving software and hardware architectures with respect to energy-

efficiency. These improvements are grouped into hardware, software, and cross-cutting approaches in the table below:

Technology/approach	Current adoption	Expected adoption in 2030	Scale of energy saving potentials
<b>Hardware</b>			
Low-power processors	● ●	● ● ●	● ● ● ●
AI accelerators	● ● ●	● ● ● ●	● ●
Task-optimised hybrid processors	● ●	● ● ●	● ●
Photonic integrated circuits	●	● ●	● ● ●
Energy-efficient memory and storage	● ● ●	● ● ● ●	● ●
Memory proximity	● ●	● ● ●	● ●
Innovative cooling technologies	● ●	● ● ● ●	● ●
<b>Software</b>			
Energy-efficient algorithms	● ●	● ● ● ●	● ● ● ●
Task-specific models	● ●	● ● ● ●	● ● ● ●
Model and code optimisation	● ●	● ● ●	● ● ●
<b>Cross-cutting</b>			
Codesign of software/hardware	● ●	● ● ●	● ●
Edge computing	● ●	● ● ●	● ● ●
Virtualisation	● ● ● ●	● ● ● ●	● ●
Intelligent energy management	● ● ●	● ● ● ●	● ●
Quantum computing	●	●	● ● ●
Neuromorphic computing	●	● ●	● ● ● ●

● = A greater number of dots indicates a higher scale

**Figure 13. Current and potential 2030 energy savings in data centers from key technologies and approaches.** Source: IEA, 2025

## Hardware efficiency

Advances in chip and system design can substantially reduce the energy intensity of AI computing. Low-power processors such as ARM-based CPUs and Intel Atom lines are being developed to minimize power draw for lightweight computing tasks. Energy-efficient memory and storage technologies, such as LPDDR5 RAM and NVMe SSDs, lower standby and active power use, while memory proximity innovations (e.g., high-bandwidth memory integrated with GPUs) shorten data-transfer distances and reduce latency. Photonics enables ultra-fast data movement and computation using light instead of electricity, dramatically reducing energy consumption. By integrating photonic components into AI hardware, systems can achieve higher throughput, lower latency, and improved scalability compared to traditional electronic architectures. Beyond processor and memory innovation, innovative cooling technologies play a critical role: liquid-cooling systems (direct-to-chip and immersion) can reduce energy used for thermal management by 20–30% compared to conventional air cooling.

Purpose-built for training and inference on AI models, AI accelerators like Google's TPU, Cerebras, Groq, Graphcore, Etched, Axelera's AI Metis<sup>23</sup> and Euclid's Craftwerk<sup>24</sup> are optimized to handle machine learning workloads at much higher efficiency compared to general-purpose chips. Task-optimized hybrid processors that combine multiple specialized chiplets maximize both performance and power efficiency by allocating compute resources to the most suitable cores<sup>25</sup>.

## Software efficiency

Energy-efficient algorithms cut unnecessary computation through techniques such as pruning, quantization, and distillation, directly lowering the number of floating-point operations (FLOPs) and energy required for training and inference; examples include Mamba<sup>26</sup> and xLSTM<sup>27</sup>. Model and code optimization refines existing model

architectures and algorithms, including compiler-level improvements, kernel fusion, memory access tuning, and hardware-aware scheduling, to ensure that software runs smoothly on available infrastructure, minimizing waste and maximizing throughput. DeepSeek is an example of energy-efficient algorithms and model/code optimization in action, employing a Mixture of Experts (MoE) architecture and using efficient attention mechanisms to reduce memory and compute cost.

Furthermore, task-specific models avoid the overhead of massive general-purpose systems by tailoring smaller, specialized architectures to particular domains, offering strong performance at a fraction of the computational cost. Models produced by Sakana AI<sup>28</sup>, for example via evolutionary model merging techniques, and their other biologically inspired architectures like CTM) are best understood as combining energy-efficient algorithmic strategies with task-directed specialization and model optimization.

Companies developing software energy-efficiency solutions can present an attractive investment opportunity because their costs of development are relatively low and the potential performance gains are high, yet their innovations are often difficult to protect due to the ease of replication and limited defensibility of software.

## Cross-cutting efficiency

Co-designing software and hardware aligns algorithmic structures with the physical characteristics of processors, to maximize performance and energy efficiency which reduces redundant computations and data transfers. From a market competition perspective, this strategy allows to combine the defensibility of hardware with the versatility of software. Virtualization allows multiple virtual machines to operate on a single server, optimizing hardware utilization and reducing the total number of physical machines required, while intelligent energy management uses AI to dynamically allocate workloads, regulate cooling systems, and optimize data center operations in real time.

---

23. <https://axelera.ai/>

24. <https://euclid.ai/>

25. <https://www.invest-nl.nl/nl/kennis-en-publicaties/semicon-deep-dive-setting-new-ambitions>

26. <https://www.ibm.com/think/topics/mamba-model>

27. [https://huggingface.co/docs/transformers/en/model\\_doc/xlstm](https://huggingface.co/docs/transformers/en/model_doc/xlstm)

28. <https://sakana.ai/>

## Edge AI

Edge computing improves efficiency by processing data closer to where it is generated, lowering the energy costs of transmitting large datasets to centralized data centers. This is particularly relevant for AI inference on small, energy-efficient devices or local edge servers, which can handle tasks like image recognition or predictive maintenance without constant cloud interaction, and where sensitive data is preferably used and stored locally.

## Quantum

Quantum advantage promises outperforming of classical computation for complex optimization and simulation problems. Quantum hardware faces challenges such as qubit error correction, and scalability as coherence decays exponentially. Existing quantum-machine learning algorithms and applications are limited in number. The impact and adoption timeframe where AI can benefit from quantum is forecast beyond the 2030 horizon of this report.

## Neuromorphic computing

Neuromorphic computing mimics the structure and functioning of the human brain to enable more efficient, adaptive, and energy-saving information processing, which could reshape AI's energy footprint across both cloud and edge infrastructures. Early-stage but promising bio-inspired technologies that are emerging as contenders for future AI-hardware include analog computing<sup>29</sup>, thermodynamic computing, reservoir computing, oscillatory and wave-based computing, and biological computing such as physical neurons and DNA-data storage. Photonics is a candidate hardware paradigm for some of these principles. Due to potential co-evolution of bio-inspired hardware and models, and potential for extreme energy-efficiency, in-memory compute, local learning enabling new forms of data privacy, and continuous interaction through event-driven sensing, the prospects of neuromorphic computing look promising.

## Efficiency, experience and expressivity

The elements of the currently dominant AI-paradigm - GPUs, tensors, backpropagation and transformers - go well together. Transformers are good for tensor-based calculations on GPUs, and their computational graphs are shallow which is good for backpropagation. They align well with next-token prediction and language data.

Ultimately though, human language connects a high-dimensional, parallel mental world with a high-dimensional, parallel world, while language itself is low-dimensional and sequential. For experiential AI, which is active, parallel, sensory and exists in-time, next token-prediction may break down: experience happens simultaneously, and the scale of tokens is not pre-defined.

There may be better candidate hardware than GPUs for processing experiential AI. Though GPUs may remain a core tenet in AI for many years, a hardware breakthrough could upend the GPU-dominated AI chip market. While digital processors such as GPUs can emulate analog and quantum processes, this tends to be highly inefficient due to the sequential nature of digital computation. Such a breakthrough could come from substantial energy-efficiency savings for general use cases such as tensor computation, inference or edge AI; or it could come from new hardware expressivity enabled by spike-dependent plasticity, thermal interaction, wave-based oscillations or other non-digital forms of computing that ushers in a new era of AI beyond the statistical machine learning paradigm.

---

29. <https://innatera.com/>

## 4. Models

The ways in which foundation models are trained, finetuned and inferenced are rapidly developing. This chapter highlights some key developments. **Model training scaling laws, where performance improves as a power law with increases in model size, dataset size, and compute, remains a foundational tool in planning large-scale model training<sup>30</sup>. Scaling laws also show that model capability scales not just with model size but with data quantity and data quality. Scaling laws help derisk experiments, because by observing smaller models one can forecast the returns of much larger ones.**

### Post-training

Post-training is commonly done with supervised instruction-tuning and alignment techniques such as RLHF and RLAIIF, parameter-efficient approaches such as LoRA, or training on a domain- or task-specific smaller dataset. It is seeing more granular and dynamic adaptation techniques. Reinforcement fine-tuning targeted at specific reasoning tasks (where correctness can be rewarded) is gaining traction. Prompt tuning, LoRA-style parameter-efficient tuning, and dynamic prompt optimization (adapting prompts per task) are being combined with feedback loops, including human ranking or automatic metrics, to continuously refine responses. Post-training adaptation will likely become more elastic and continuous. Instead of one-shot tuning, models may maintain live adapters that fine-tune themselves from deployed usage, effectively bridging online learning with safety constraints.

### Reasoning and the inference scaling paradigm

Inference is the ongoing service workload where the tuned model produces outputs to queries. Inference is the process of running a trained and tuned model to generate predictions or decisions based on new, unseen data such as user queries. While training used to dominate attention, inference is now an increasingly critical battleground. Inference cost per token will continue to fall due to hardware optimizations, quantization, pruning, and other innovations, but total serving costs keep rising as it scales with users and usage rather than with model release cycles, and reasoning-style queries impose multiple forward passes or internal search. AI training vs inference workload is expected to shift from 20-80 in 2023 to 15-85 in 2028<sup>31</sup>. AI inference applications are also increasingly moving to the edge, such as edge data centers and end-user devices, with AI central vs edge workload expected to shift from 95-5 in 2023 to 50-50 in 2028<sup>32</sup>.

A promising post-training direction is distillation of reasoning modalities: training smaller models to mimic multi-pass reasoning outputs. Early experiments show that with proper distillation and constraints, much of the performance of expensive reasoning LLMs can be compressed and deployed efficiently. In the coming years, hybrid models (reasoning-augmented but distilled) will likely be commonplace in production systems, i.e. occasional use of heavier reasoning core, typical use of a fast distilled model.

Beyond static inference, the idea of test-time scaling, where more compute is applied at inference time adaptively, is gaining traction. Research on compute-optimal inference suggests that for certain tasks, it's better to allocate extra inference compute

30. <https://cameronrwolfe.substack.com/p/llm-scaling-laws>

31. source: European Commission, Scheider Electric

32. source: European Commission, Scheider Electric

than to scale model size further<sup>33</sup>. As a corollary, chip roadmaps increasingly optimize for inference throughput and low-latency operations, not just training.

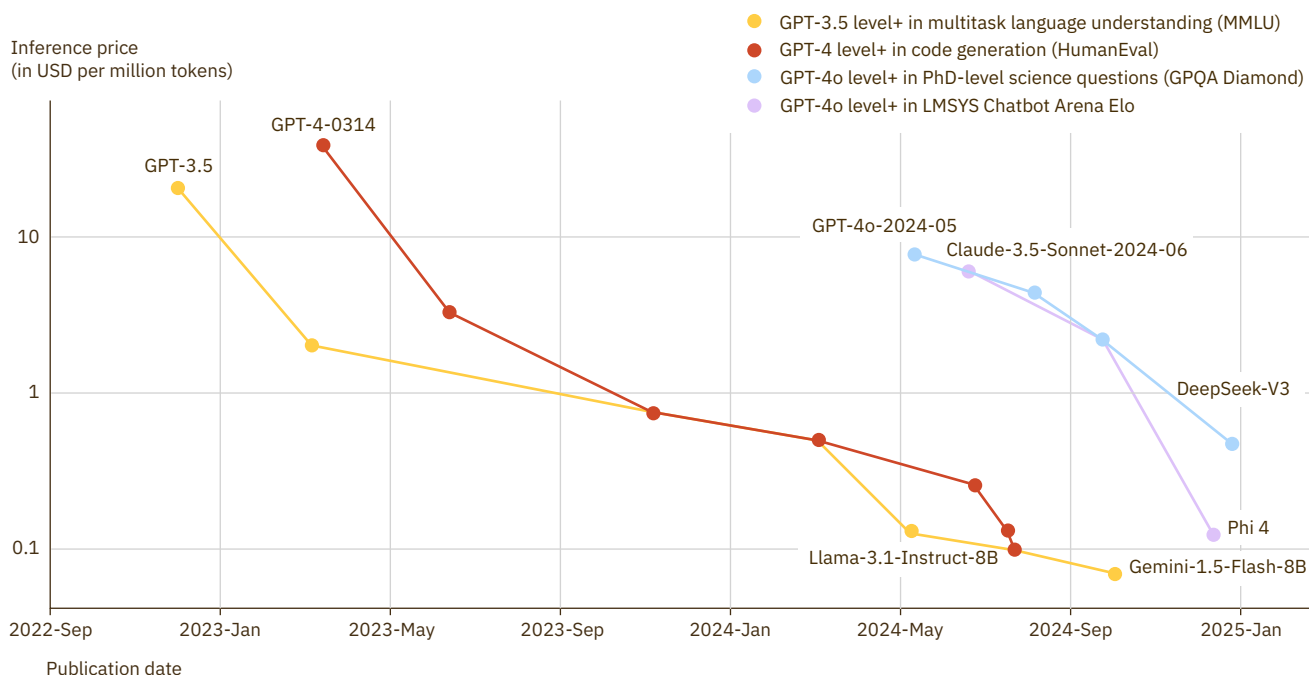


Figure 14. AI inference price for customers, per 1 million tokens. Source: AI Index Report, 2025

## Open source

Open source and open weights models are key to AI transparency and accessibility. Open weights refer to model parameters being publicly available, which allows for independent evaluation, deployment, and modification of AI models. Open weight models, such as Mistral’s models and Meta’s Llama, enable secure and private deployment: since these models can run locally without depending on third-party APIs, users retain control over their data and avoid the risks associated with remote inference, an especially critical factor in privacy-sensitive or security-critical applications.

However, the availability of open weights alone does not make a model open in the traditional software sense. Without open training code, training datasets, and comprehensive system or model cards documenting preprocessing, architecture, and hyperparameters, replication becomes prohibitively expensive and less reliable. Licensing terms also vary widely—from permissive

ones like MIT, used for DeepSeek-R1, which allow unrestricted commercial use and derivative works, to more restrictive ones like those applied to Meta’s Llama models.

Though open-source models typically lag six to twelve months behind closed-source frontier models, they remain a viable approach in domains where transparency, community-driven innovation, and distributed ownership are valued alongside model performance, or where head-on competition on compute and data are simply not an option. With current top three open-source LLM being DeepSeek, Minimax, and Qwen (Alibaba), China is spearheading this approach, and Europe would do well to follow this strategy for LLMs where it can’t participate in competition at scale but also can’t rely on robust foreign access.

33. <https://blogs.nvidia.com/blog/ai-scaling-laws/>

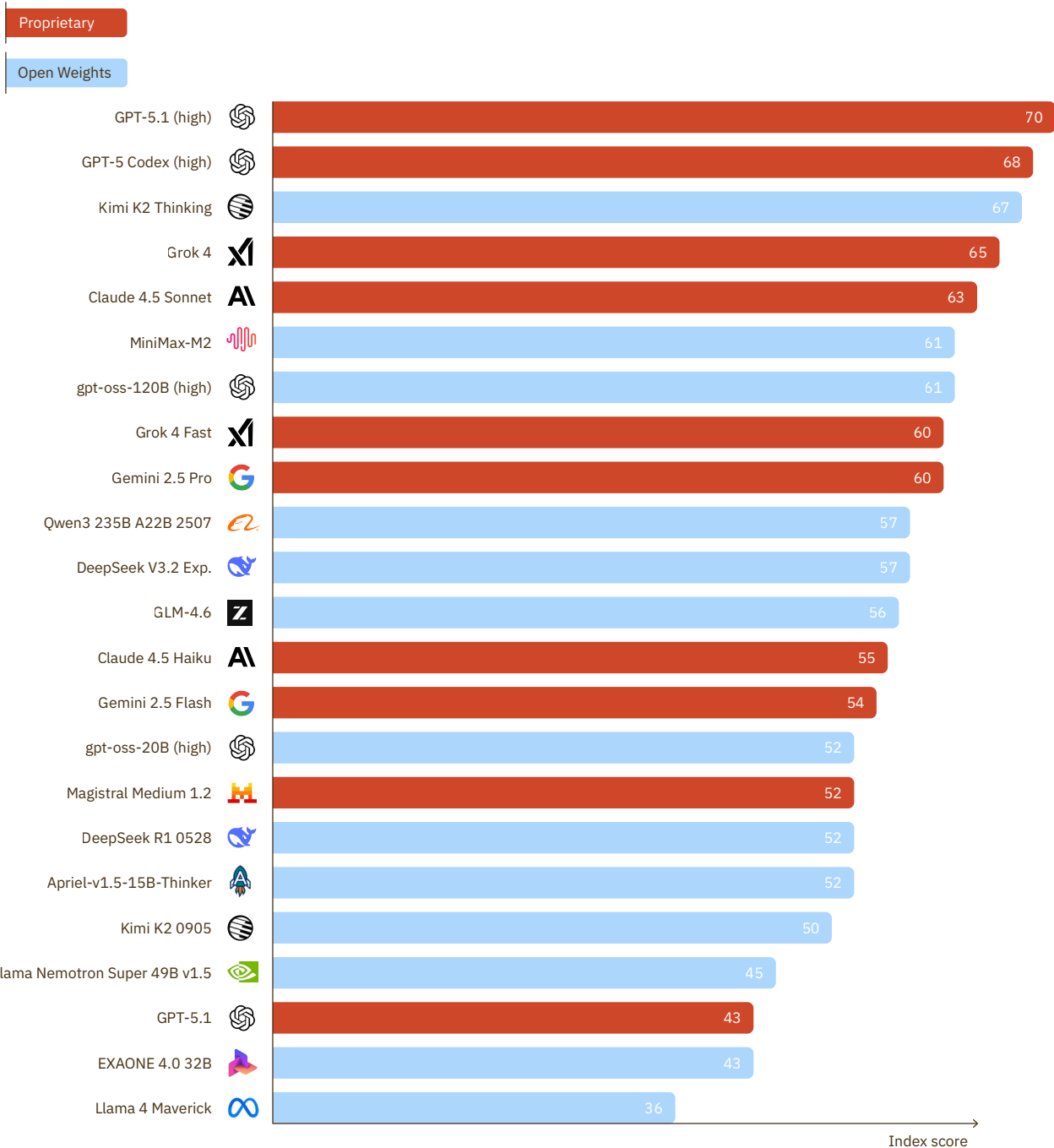


Figure 15. Evaluation index of proprietary vs open weights models. Source: Artificial Analysis, November 18th 2025, see Annex II

## Continual and online learning

In continual learning, models incrementally acquire new knowledge while retaining previously learned information, using mechanisms such as modular architectures, replay strategies, and parameter isolation to reduce catastrophic forgetting. Online learning updates model parameters sequentially as new data arrives, enabling rapid adaptation to non-stationary environments and fine-grained control over model drift. As these approaches mature, they point toward a future in which training

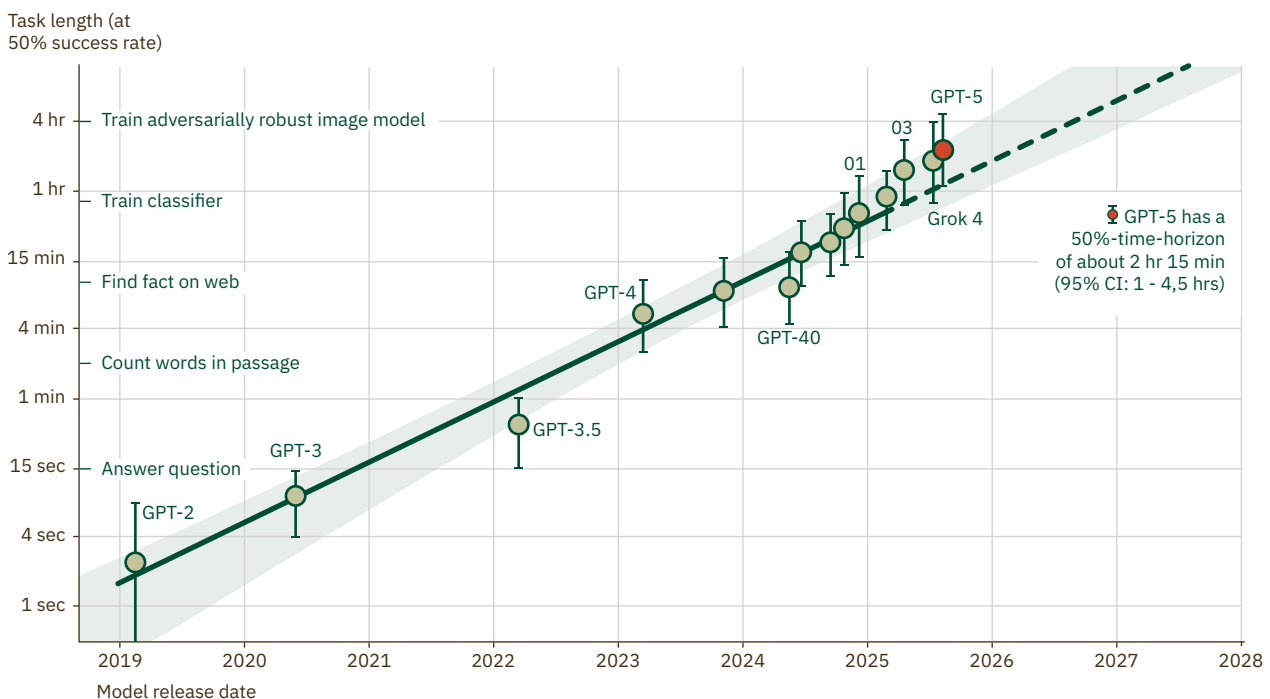
and inference become increasingly intertwined: models will update themselves during deployment, continuously refining parameters based on real-time observations rather than relying on fixed offline training cycles. In combination with a shift to open-source models, this integration between training and inference offers opportunities as future models may not be deployed standalone off-the-shelf but rather connected to trusted base models.

## Small models

Despite the theoretical promises of scaling laws and the astonishing breakthroughs made with large foundation models in recent years, every exponential curve becomes a sigmoid at some point. The scaling frontier is under mounting tension, and some voices in the field argue that blindly scaling is no longer the central path forward<sup>34</sup>, warning of an end to low-hanging fruits for AI growth, or that AI is a bubble about to burst. In response, researchers are exploring more efficient scaling strategies such as sparse architectures, mixture-of-experts (MoE) models that only route parts of the network per token, and pruning-aware training. Future monolithic models will likely be constructed as ensembles, mixtures of specialized modules, or as a superset of experts with only partial activation. Furthermore, many workflows are narrow, repetitive and format-bound, so small models are often sufficient and less costly. This, combined with inference-aware pretraining, suggests that the sweet spot of scale might change from “bigger is better” to “bigger where it matters, lean where it doesn’t”.

## Agentic AI

Agents are automation engines for information work, that can do planning, retrieving, synthesizing information, use tools and act across digital ecosystems with limited human intervention<sup>35</sup>. The rise of agentic AI requires new benchmarks for evaluating usefulness: rather than measuring only response accuracy or fluency, models are evaluated on planning coherence, tool use, feedback loops, trustworthiness and many other capabilities. Comparing their work to human capabilities misses the point entirely. As agent capabilities mature, automation of information work won’t limit itself to the human pace of work, decision making or question answering. Progressive task length combined with increased speed of execution will minimize costs and time required to execute entire information workflows, turning what used to be human-intensive work into composable, agent-managed processes. This allows AI to augment and automate existing human workflows, and to automate new workflows that are not economically viable for humans to carry out.



**Figure 16. Task length benchmarks offer another perspective of looking at capability improvement, showing that progress over the last years has followed a surprisingly constant trend.** Source: METR, 2025

34. <https://www.businessinsider.com/meta-yann-lecun-scaling-ai-wont-make-it-smarter-2025-4>  
 35. <https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-an-ai-agent>

# Artificial General Intelligence

Critics of agentic AI caution against a blind belief in agentic AI and warn that they merely amplify illusions of intelligence. Results may be attributed to overfitting on benchmarks, and many pilots still fail to produce meaningful impact in a business context<sup>36</sup>. Other critics caution against threats and disruptions caused by agents with superhuman abilities. Key to these speculations is to what extent agents will be capable of recursive self-improvement, where they iteratively produce more capable successors at an accelerating pace. Scaling models is a form of internal self-improvement, as new capabilities emerge from scale. Similarly, agents can learn from other agents and refine objectives, for example using reinforcement learning to generate new reward functions. AI is also a tool for humans to build better AI systems, accelerating development through AI-assisted coding and AI-generated chip designs. Scenarios where agents recursively self-improve without any form of human intervention could potentially lead to dramatic acceleration, though such scenarios tend to presuppose a strong functionalist view of

intelligence being fully reducible to computable processes, and oversimplify the human experience. Ultimately, AI as we know it may be key to creating a future systems technology capable of autonomously understanding, learning, and performing any intellectual task that humans can.

## Integration

AI model training is only one part of the equation; applications deliver value when infrastructure, data, models and user interactions are integrated well. Approaches such as retrieval-augmented generation (RAG, integrating large language models with vector databases), neuro-symbolic systems (combining neural networks with symbolic reasoning), and agentic models and workflows (LLMs with memory, planning, and tool use) reflect this. Effective integration also depends on building trust through governance, orchestration, security, and compliance: elements that ensure AI systems operate transparently and responsibly.

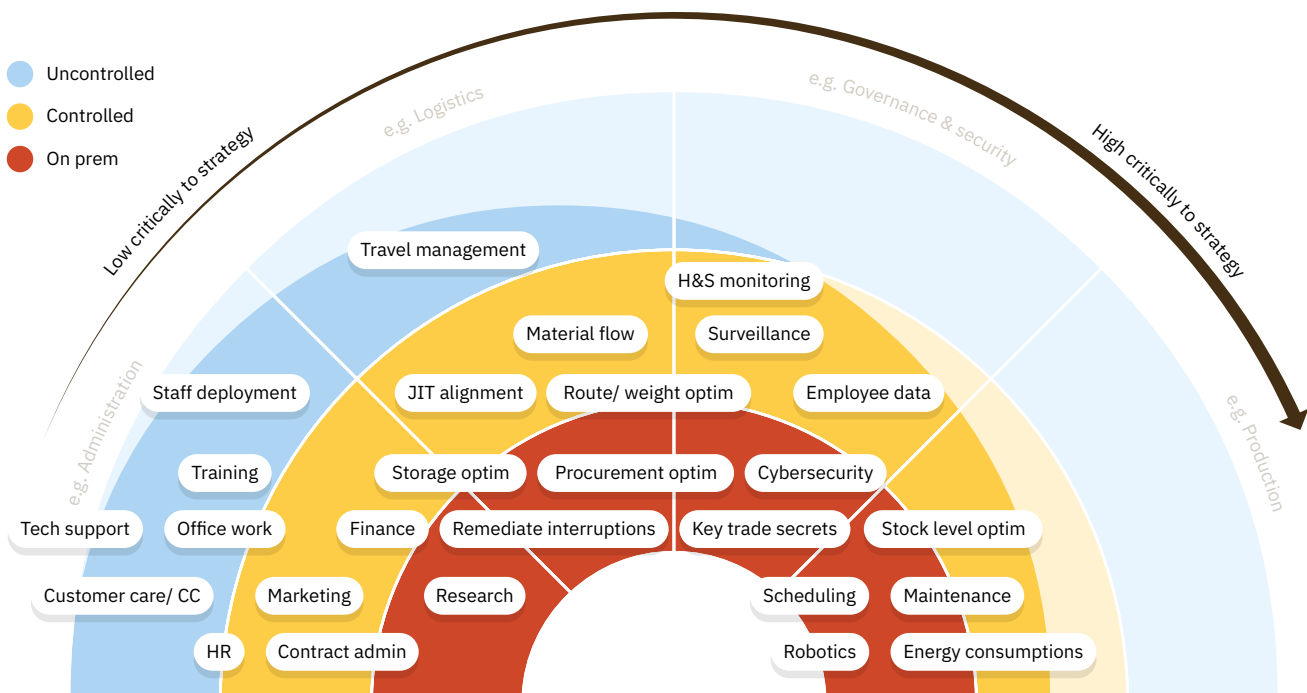


Figure 17. Criticality and control of functions powered by AI. Courtesy of Arthur D. Little

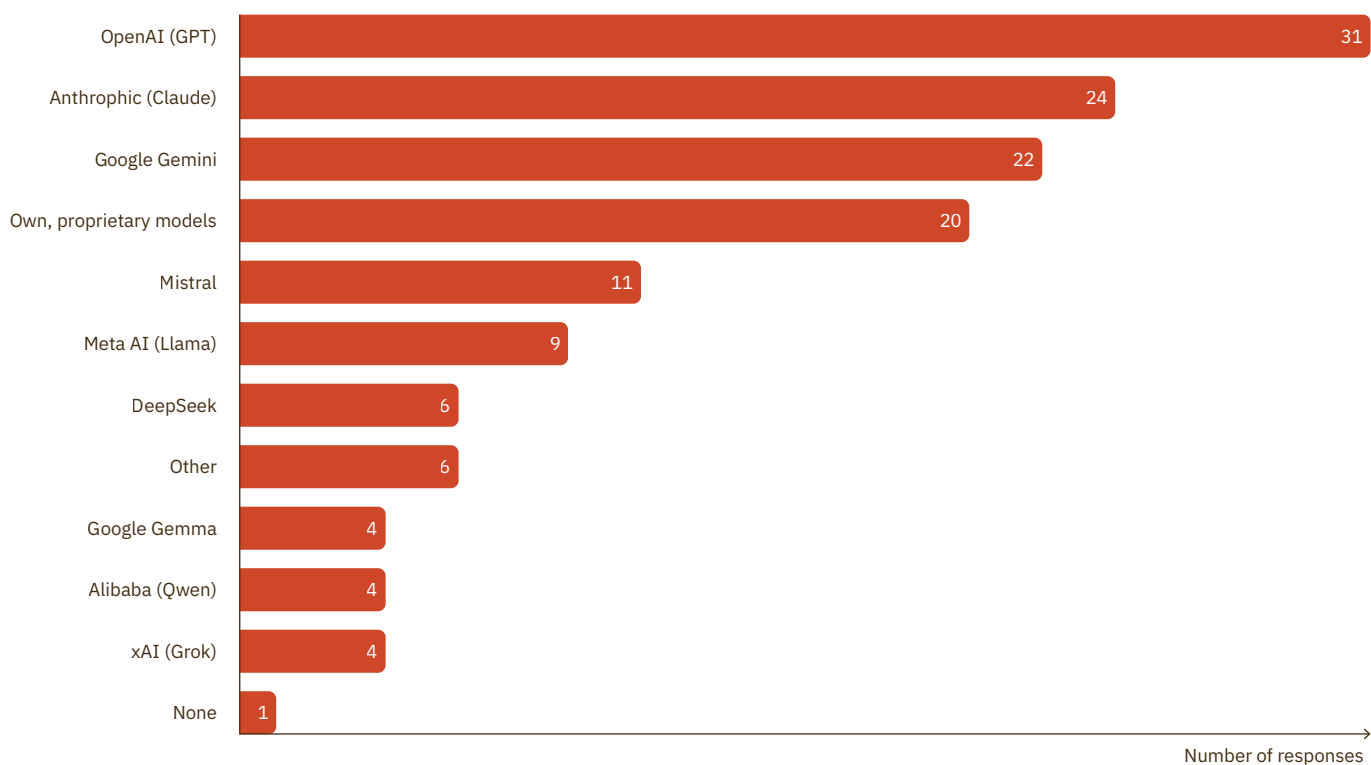
36. State of AI in Business (MIT, 2025)

The deployment environments of AI-powered organizational functions differ in terms of control. Data sensitivity and operational criticality decide to what extent AI workloads should be hosted on premise e.g. using open weight models or in controlled environments e.g. through APIs with trusted third parties. For example, applications in contract administration, HR, marketing, and scheduling on the less controlled side, while more sensitive areas like cybersecurity, robotics, and energy consumption typically require controlled or on-prem environments.

Major platforms incumbents like Google, AWS, and Microsoft achieve integration by aligning components across horizontally and vertically across the technology stack, while other emerging contenders such as OpenAI, Perplexity, and Mistral seek collaboration for vertical integration, as are Nvidia (moving into robotics) and Grok (potential integration into X and Tesla).

For Europe and the Netherlands, this trend raises concerns about platform dependency and switching costs, as many AI capabilities are embedded in U.S.-based clouds and APIs. Since model cost and performance are not the only deciding factors, issues like privacy, control, and regulatory compliance increasingly shape adoption. The likely future will be neither fully vertical nor entirely deverticalized but a mixed landscape that varies with sectoral and use case requirements.

Resource constraints, not limitless compute, may be the key to AI innovation for Europe. With a business model that focuses on AI deployment and integration into existing platforms and processes, European companies can differentiate in terms of data security, derisking for foreign dependence, and finding foundation models fit business context, rather than relying on output produced by the largest model<sup>37</sup>. This strategy benefits from energy efficient AI hardware, software, and co-design. It is employed by companies such as Mistral<sup>38</sup> and Sakana<sup>39</sup>.



**Figure 18. Model providers for Dutch AI startups.** Source: AI Deepdive Survey, October 2025; see Annex II

37. c.f. Christophe Fouquet in Buitenhof on ASML's investment in Mistral, <https://www.vpro.nl/buitenhof/artikelen/buitenhof-16-november-2025>

38. <https://mistral.ai/>

39. <https://sakana.ai/>

# 5. Market

To forecast the value and impact of AI in the coming years, we look at how AI will be used, the scale of private enterprise and venture capital investment, and how governments shaping markets through innovation policy and regulation.

## AI usage

AI usage spans both organizational and personal contexts, with distinct patterns. On the enterprise side, adoption is strongest in marketing and

sales, product development, and IT departments, particularly for technology, professional services, advanced industries, and media and telecom.

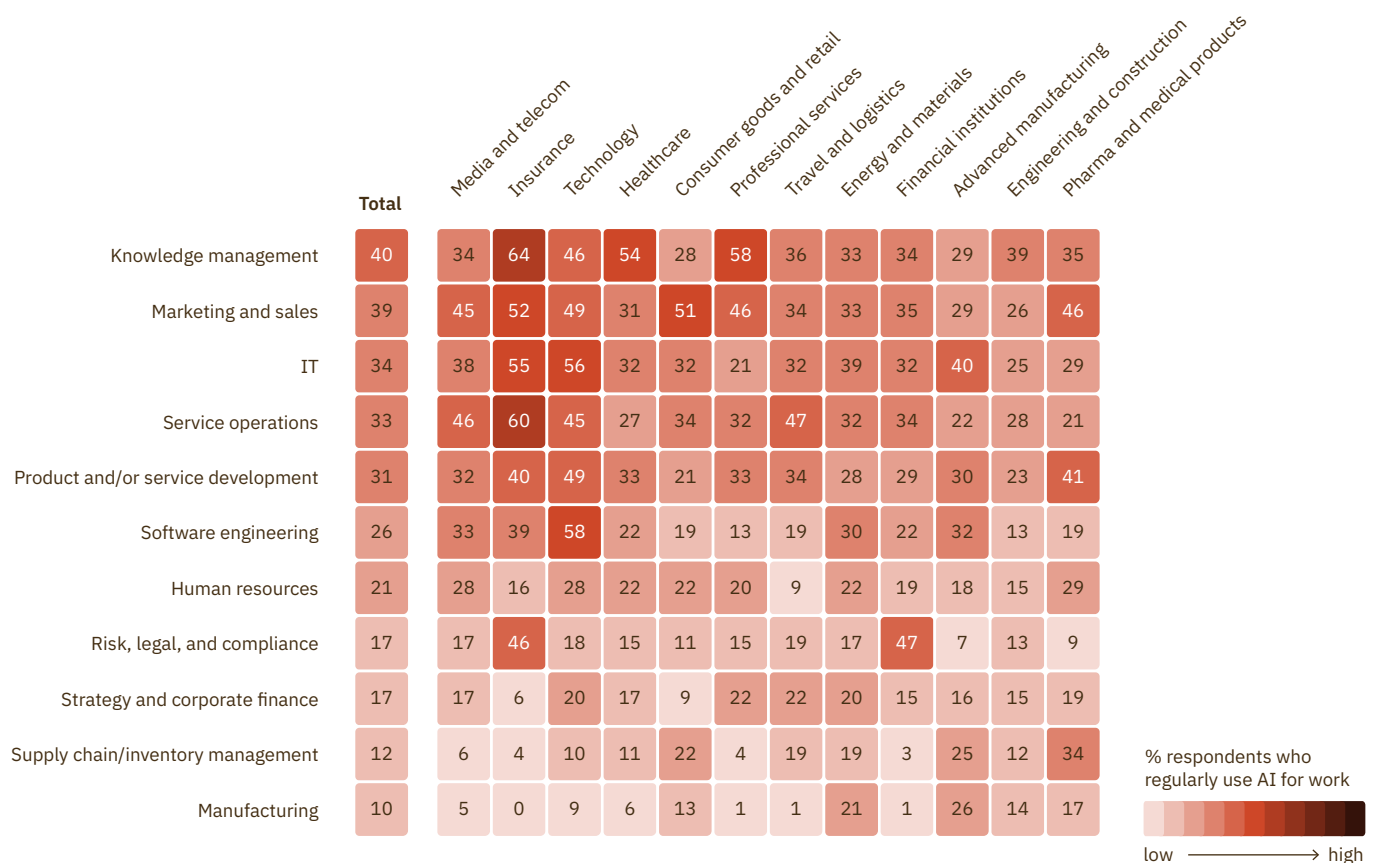


Figure 19. Business functions in which respondents' organizations are regularly using AI, by industry (% respondents) Source: The State of AI in 2025, QuantumBlack, 2025

## Regulation

The European AI Act entered into force in August 2024 and will be applied in phases. Since February 2025 the unacceptable-risk AI practices are already banned. By August 2026 high-risk AI systems, those used in critical infrastructure, education, employment, migration, and other application areas, must comply with the full set of obligations including risk-management, quality-management, human-oversight, technical documentation and logging.

The AI Act is paradoxically considered both a barrier to innovation as well as an enabler. There are concerns that compliance and administrative burden will lower the competitiveness of European AI developers, suppliers and deployers. The AI Act 2.0 will likely address some of these concerns, lessen or delay administrative burdens on companies, and pull more regulatory oversight mandate into the Brussel's AI Office rather than national supervisory bodies.

For some companies, regulatory hurdles and data fragmentation can sometimes be key advantages, for example in highly regularized sectors such as healthcare and finance, as it keeps competitors at bay. For these companies, scaling markets internationally can be a risk as they face less competitive advantage from these obstacles in other markets.

## Global market

Market sizing studies for AI tend to be unreliable because the technology and its applications evolve rapidly and the scope of what is considered AI varies, from language models alone to systems that combine language, multimedia, and code, or even broader integrations. This makes consistent measurement challenging. A 2023 McKinsey Global Institute study estimates that new generative AI use cases will yield a US \$2.6-4.4 trillion economic impact, on top of the US \$11-17.7 trillion impact from advanced analytics, traditional machine learning and deep learning. The economic impact of all worker productivity, including new use cases, is estimated to yield an ~35-70% increment, yielding a total AI economic potential of US \$17.1-25.6 trillion globally.

The US \$2.6-4.4 trillion impact from new generative AI use cases can be broken down further per sectors and business function. Notable is the economic impact of software engineering in high-tech, risk and legal use cases in banking, and product R&D and pharmaceuticals and medical products.

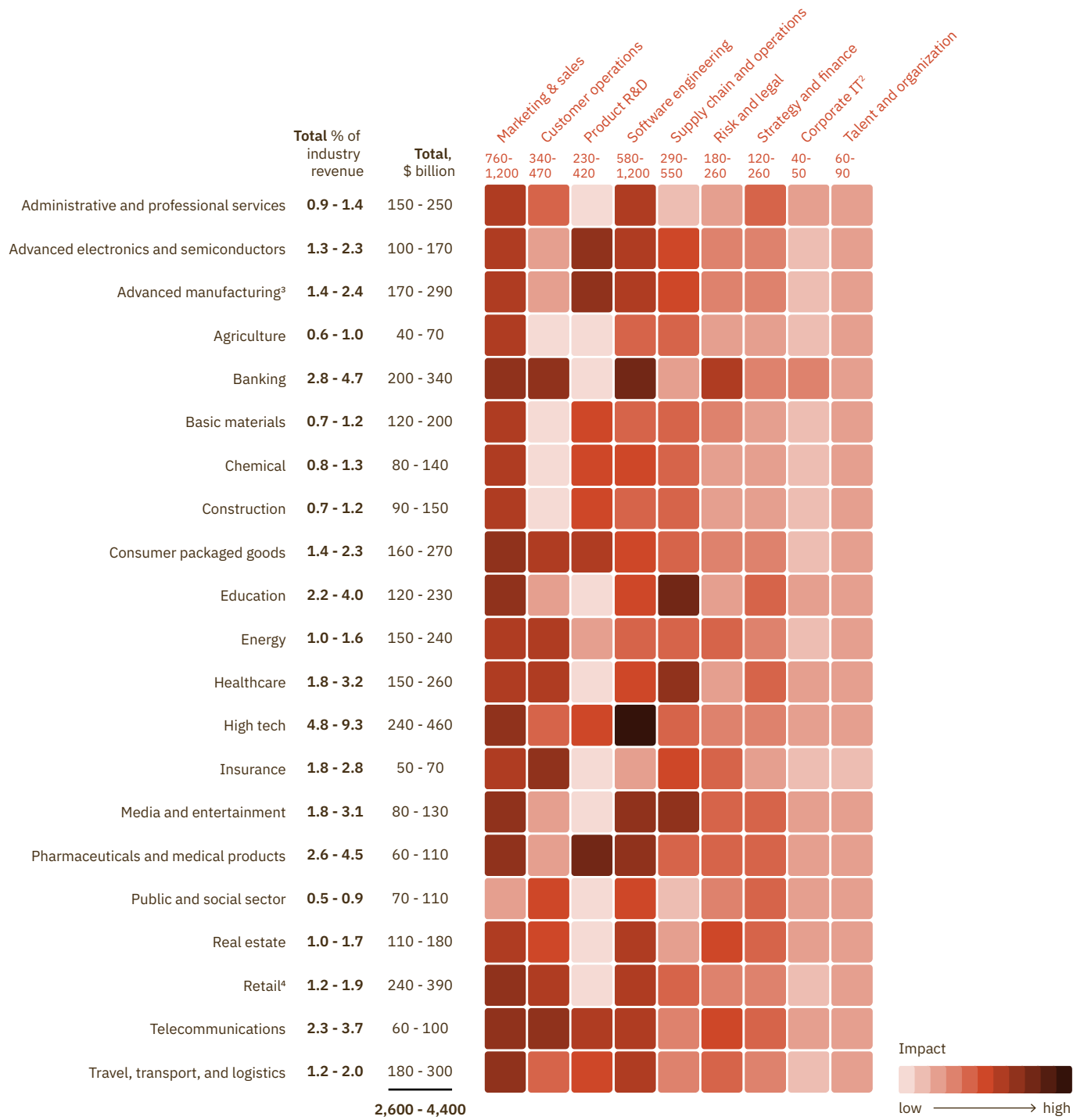


Figure 20. Generative AI productivity impact by business functions. Source: The Economic Potential of Generative AI, McKinsey Global Institute, 2023

## Private Investments

AI's technological momentum has been matched and magnified by extraordinary levels of investment. In 2024, global private AI investment reached US \$109 billion, including US \$33.9 billion directed specifically toward generative AI. In 2025, Microsoft, Meta, Amazon, and Alphabet are

expected to spend nearly \$400 billion collectively on AI-focused capital expenditures. Nvidia forecasts US \$ 3-4 trillion in global AI infrastructure spending by 2030<sup>40</sup>. Morgan Stanley projects US \$2.9 trillion in AI-related investment between 2025 and 2028.

40. <https://www.reuters.com/business/nvidia-ceo-says-ai-boom-far-over-after-tepid-sales-forecast-2025-08-28/>

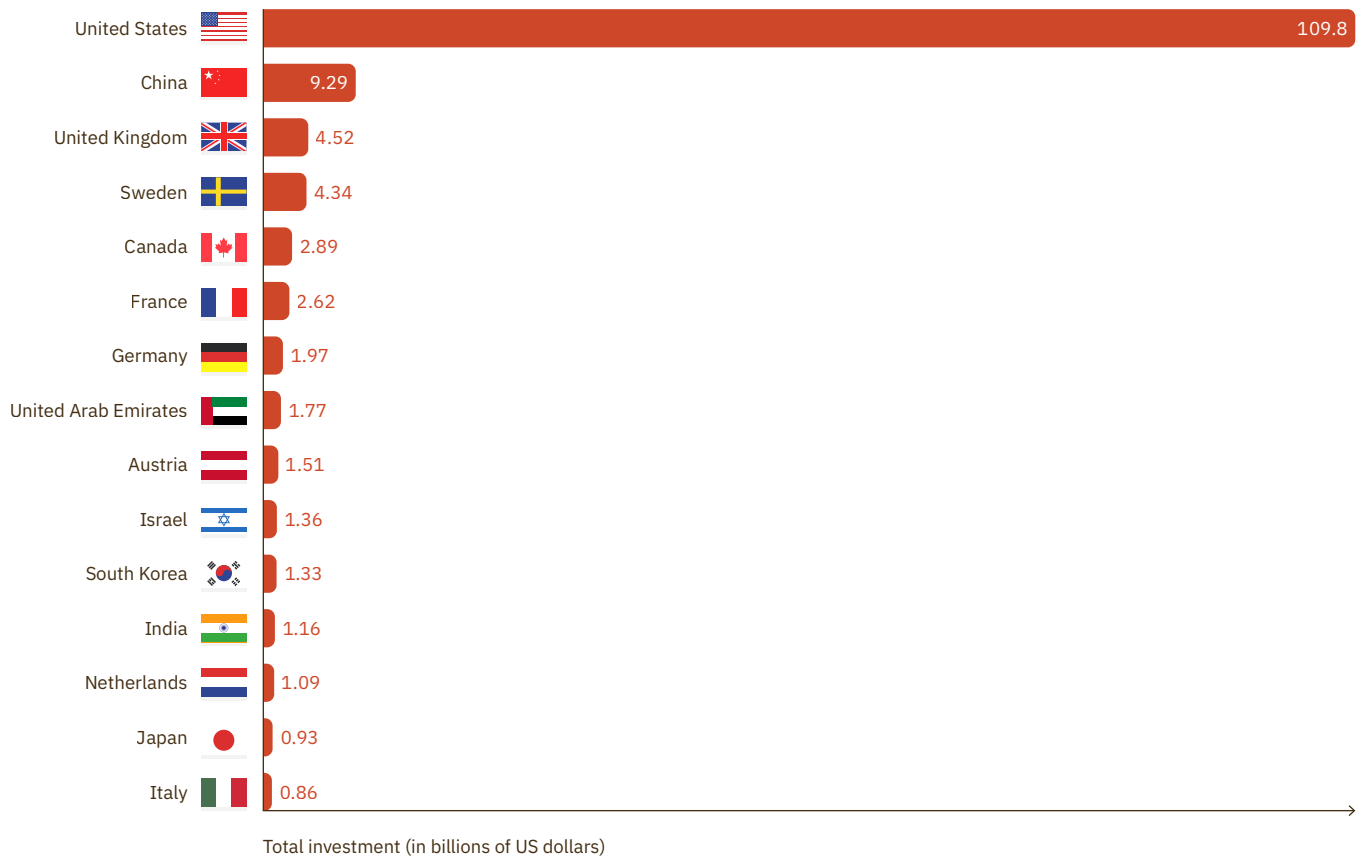


Figure 21. Global private investment in AI by geographic area, 2024. Source: Quid, AI Index Report, 2025

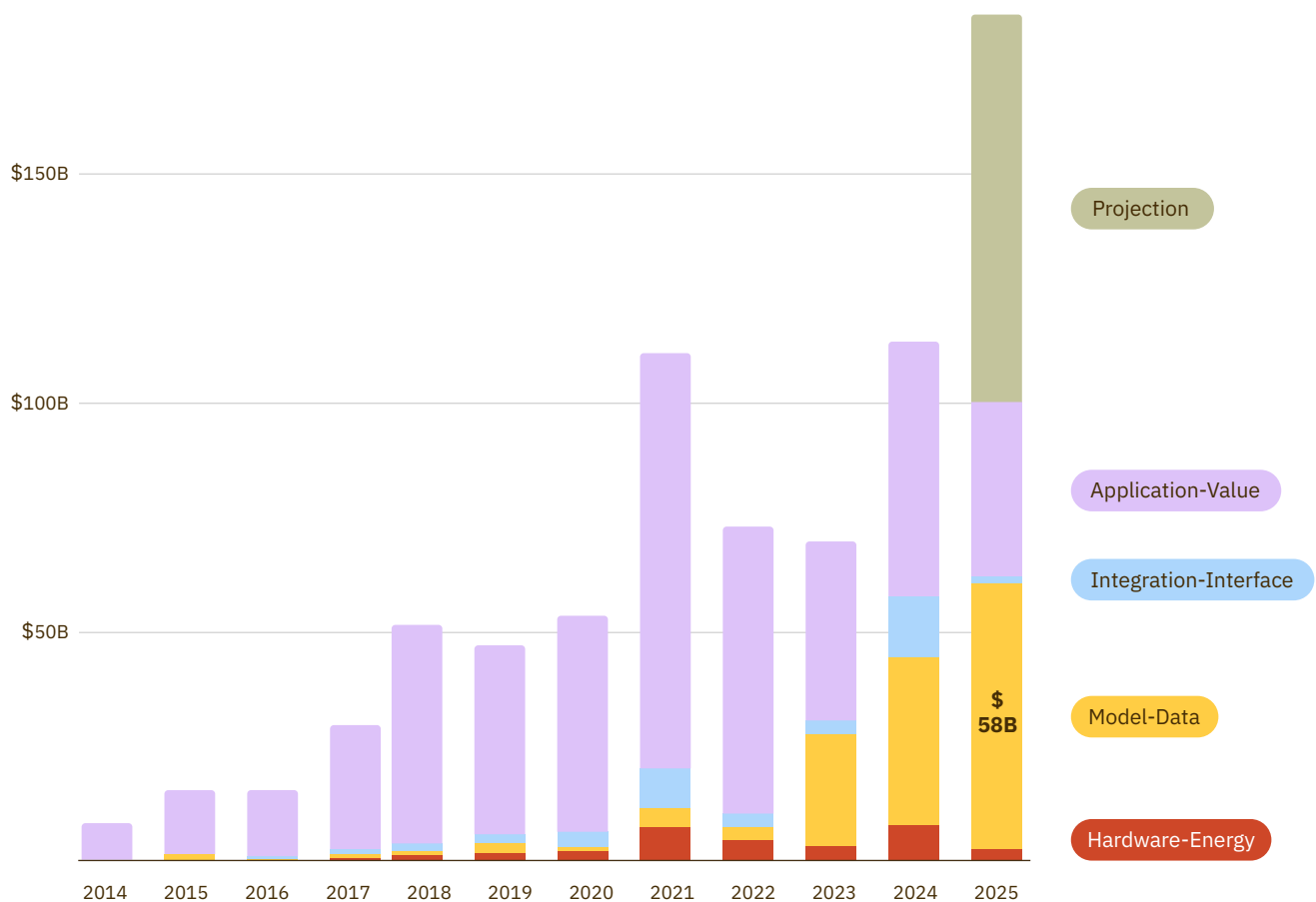


Figure 22. Global AI VC investment by layer. The Model-Data layer drove most of the growth in the last two years Source: Dealroom, 2025

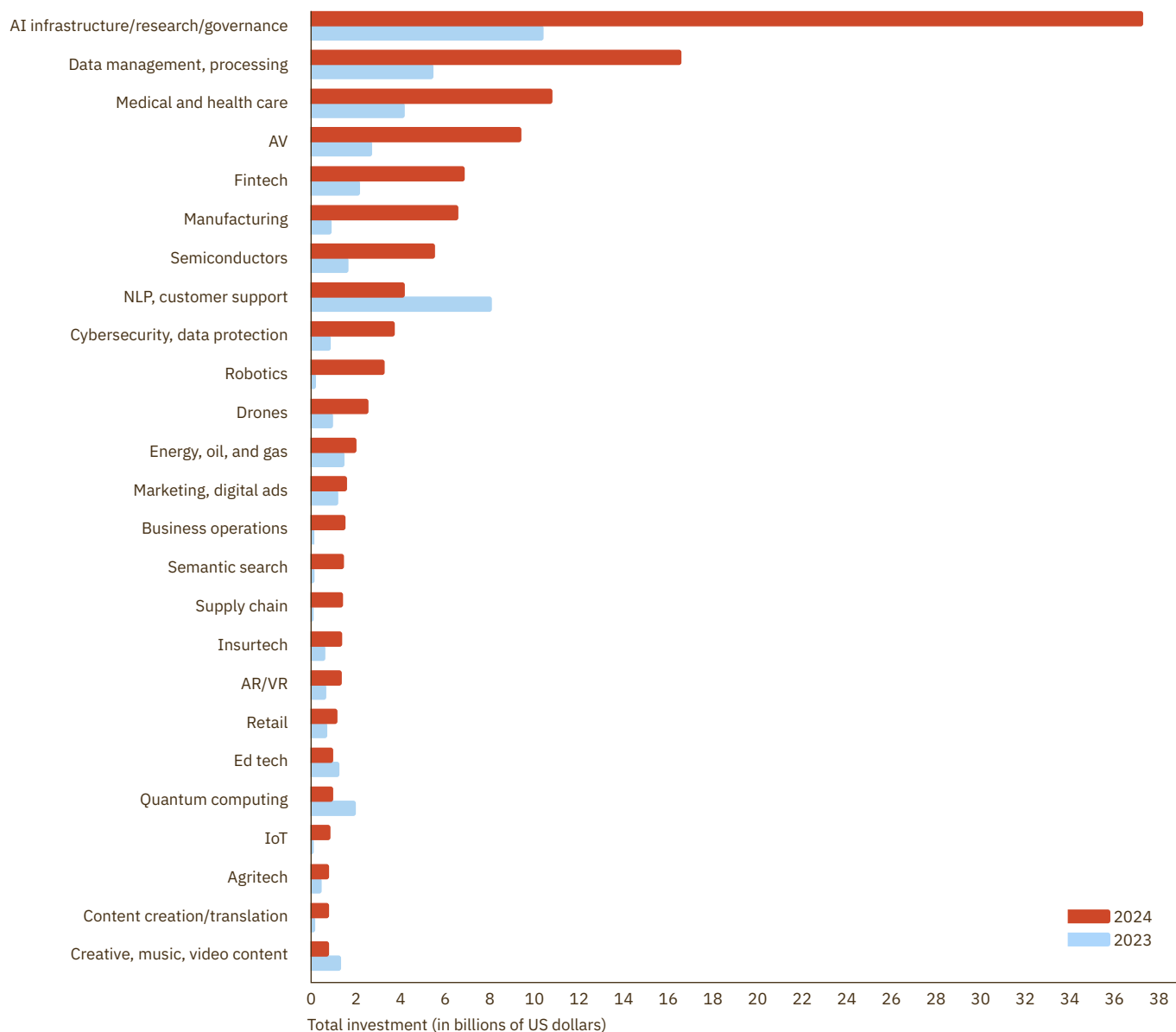


Figure 23. Global private investment in AI by focus area, 2023 vs 2024. Source: Quid, AI Index Report, 2025

## European Market

Reports generally estimate a 30% CAGR from 2025-2030 driven by continued AI adoption. Given a baseline spend of €46 billion in 2024<sup>41</sup>, we can forecast the annual gross value created in Europe<sup>42</sup>. Using industry averages, we calculate that enterprises and public sector intend to create ~€ 500 billion in value in 2030.

The InvestAI EU initiative is a multi-year program that mobilizes a total of € 200 billion in investments for AI to strengthen Europe’s AI sovereignty,

infrastructure and competitiveness. €20 Billion of this is earmarked for AI gigafactories, HPC centers for training and inference. Though Europe is stepping up, these figures are not comparable in scale to US and Chinese AI investments. Europe’s AI landscape is also fragmented across nations, with different levels of readiness, policies, and existing infrastructure; harmonizing efforts is nontrivial. Europe faces a significant structural disadvantage due to its high energy prices<sup>43</sup>. And, though built for trust, there is a growing

41. combined spend of EU-27 enterprise and public sector on AI, including hardware, platforms, API’s, software and services (source: IDC)  
 42. see annex II  
 43. Draghi (2024)

worry that Europe's regulatory approach to digital developments, including the AI Act, GDPR, DSA and DMA may hamper experimentation unless compliance support and simplification are handled well.

In this context, it makes sense for Europe to adopt a distinctive and complementary strategy rather than competing head-to-head on scale. Europe's advantage lies in its knowledge ecosystems, smaller yet efficient models, and domain-specific data strengths - particularly in scientific, industrial, and public sectors. The InvestAI initiative should be viewed in light of building resilience and sovereignty by prioritizing public, pre-competitive infrastructure, and fostering collaboration among academia, startups, and established industries.

AI is a global market. 17% of the world population lives in China, 9% in Europe, 4% in the US, 70% lives in the rest of the world. Europe has opportunity to capitalize on this market if it manages to leverage trustworthiness as a competitive advantage and focuses on applying AI for industrial and public challenges e.g. in agriculture, climate change and healthcare.

# 6. The Netherlands

The Netherlands shows moderate strength in infrastructure and operating environment but lags behind in research, development, and commercial investment.

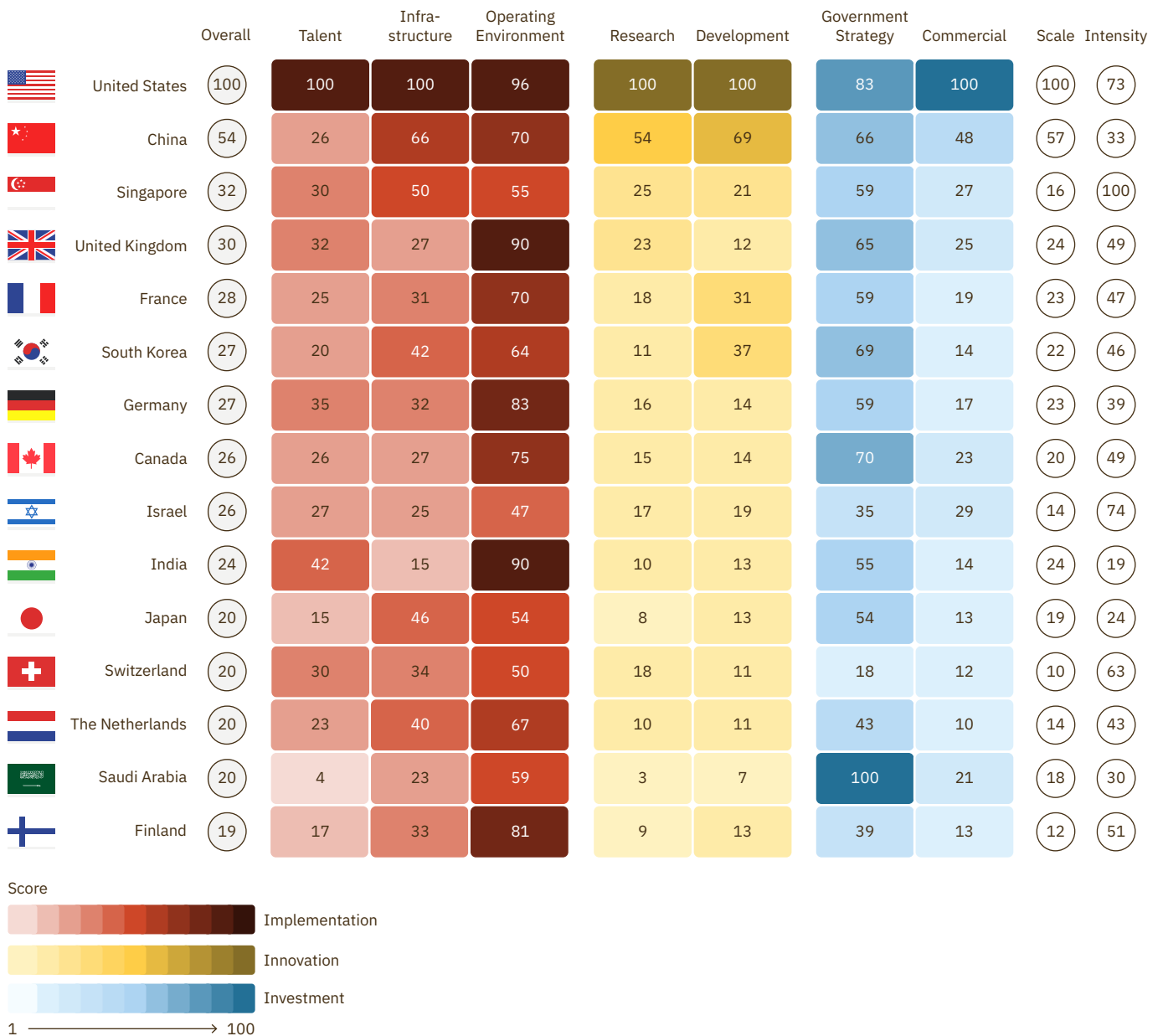


Figure 24. AI capacity ranking by country. Source: The Global AI Index, Tortoise Media, 2024

● The larger the dot, the higher the number of startups

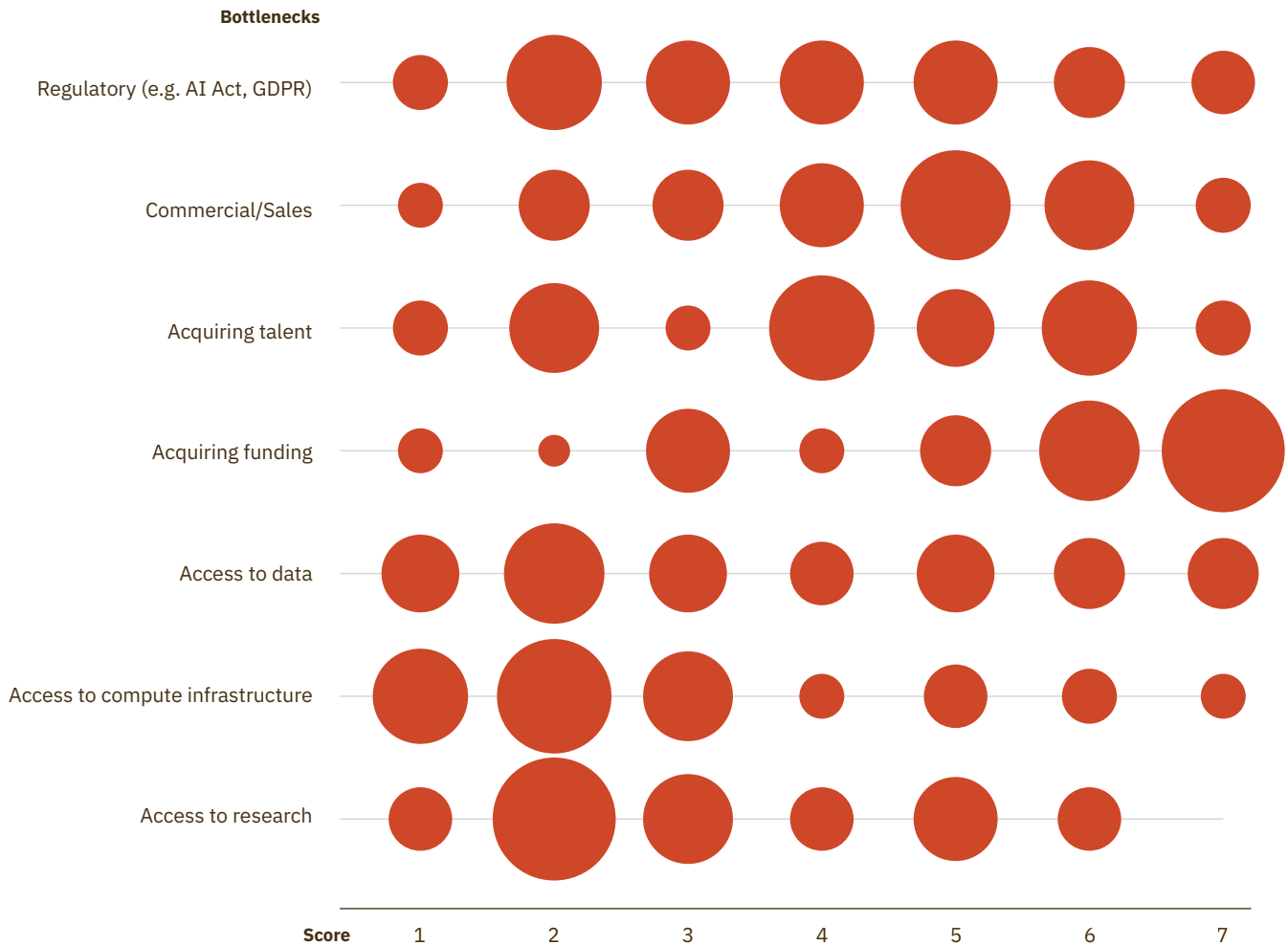


Figure 25. Growth and innovation bottlenecks for Dutch AI startups. Source: AI Deepdive Survey, October 2025

## Energy

Europe's position does not benefit from the relatively price it pays for energy: its industrial retail power price is 158% of the price the US and China pay<sup>44</sup>. Even within Europe, the Netherlands has a high price point for electricity (95 eur/MWh) compared to Germany (46 eur/MWh), France (32 eur/MWh) and Belgium (56 eur/MWh) due to relatively few tax exemptions and compensations.

## Knowledge ecosystem

All Dutch research universities host AI groups, but several stand out as international leaders. Universiteit van Amsterdam (UvA) has a world-class presence in deep learning, vision and generative models, with very high publication output in top AI conferences such as NeurIPS, ICML and CVPR. They have a key role in ELLIS<sup>45</sup>, and with AMLab, QUVA, Bosch Delta Lab, Qualcomm Lab they host excellent ICAI labs<sup>46</sup>. TU Delft is strong in embedded AI, energy systems, mobility, and human-AI interaction, and engineering-driven AI: robotics, autonomous systems, and safety. Delft has excellent industry links, for example to

44. European Commission (2024), Eurostat, EIA

45. <https://ellis.eu/>

46. <https://icai.ai/>

NLR, TNO, and ESA, and hosts another ELLIS unit. Utrecht University (UU) is known for their research in hybrid AI, knowledge representation, and human-centric AI, and ties to life sciences, mobility, energy, governance, and health.

Radboud University Nijmegen is strong in cognitive AI, fundamental machine learning, brain-inspired learning, and medical imaging, with Donders Institute giving Nijmegen an edge in neuro-AI. Nijmegen also hosts an ELLIS unit. TU Eindhoven (TU/e) is known for computational intelligence, high-tech systems, mobility, manufacturing and photonics, with ties to applied AI in industry such as Philips and ASML. Universiteit Leiden specializes in explainable AI, computational linguistics, and AI governance, and has ties with bioinformatics and medical data science at LUMC. The University of Groningen (RUG) is recognized for its strengths in machine learning, data science, and socially responsible AI, with a strong focus on natural language processing and AI applications in energy and sustainability. Vrije Universiteit Amsterdam (VU Amsterdam) is known for in hybrid AI and responsible AI systems, supported by strong interdisciplinary research across computer science, social sciences, and behavioral sciences.

Universiteit Twente has strengths in human-computer interaction, neuromorphic computing, AI ethics, and robotics. Tilburg University specializes in AI governance, law, society, and digital transformations. Maastricht University is known for AI in healthcare and federated learning. Wageningen University & Research (WUR) is a leader in AI for agriculture, food systems, and environmental sustainability, with expertise in data-driven crop management, smart farming, and ecological modelling. Erasmus University Rotterdam focuses on AI for business, economics, and society, and AI-driven decision-making in healthcare and public policy.

While Dutch universities are strong regionally and have niche AI strengths, many do not match the top tier of European heavyweights in terms of global brand, volume of elite publications, and volume of funding. They have clear thematic focus and are well-integrated in the Dutch and European ecosystem, through national AI hubs and industry partnerships, which grants good leverage for applying AI research. The Netherlands generally

hosts a good research-policy environment, relatively high openness and collaboration, and decent proximity to industry clusters.

Next to research universities, TNO plays a bridging role, applying AI to domains such as defense, energy, mobility, and industry. CWI contributes foundational breakthroughs in mathematics, algorithms, and data science that underpin AI. The universities of applied sciences (HBOs) research responsible application and pilot AI with SMEs, regional industries and governments. ICAI (Innovation Center for Artificial Intelligence) is a Dutch national network 55 public-private research labs and over 600 researchers nationwide.

The Netherlands ranks first on the global index on responsible AI (91.12), ahead of Germany (90.94), Ireland, (73.68), UK (79.62), and the United States (74.75)<sup>47</sup>. Part of this frontrunner position is due the national network of over 20 ELSA Labs (Ethical, Legal and Societal Aspects), collaborative testbeds where universities, companies, government bodies, and civil society organizations co-develop AI solutions with explicit attention to ethics, law, and societal impact.

## Talent

Frontier AI development demands highly specialized expertise. Investors interviewed consistently cite the quality of the team as the most important factor influencing their investment decisions, underscoring the need to invest strategically in top talent across key parts of the AI stack and industry sectors. The global battle for talent hinges on what countries can offer: quality of life, competitive salaries, access to compute and data infrastructure, interesting and meaningful problems, and vibrant ecosystems that enable knowledge exchange and collaboration. The Netherlands has good foundations in academia, talent, and quality of life, giving it the potential to compete internationally. Inflow of students in higher education AI studies has increased from 1968 in 2015 to 6103 in 2023<sup>48</sup>. However, its limited capacity to scale in terms of compute, capital, and market results in an underdeveloped operating environment, causing valuable talent to migrate abroad. Nearly 70% of Dutch AI startups

---

47. <https://www.global-index.ai/>

48. <https://pr-edict.nl/themarapportages/ai-en-data-science>

employ fewer than 10 people, underlining both the ecosystem’s entrepreneurial vitality and its structural scaling challenges.

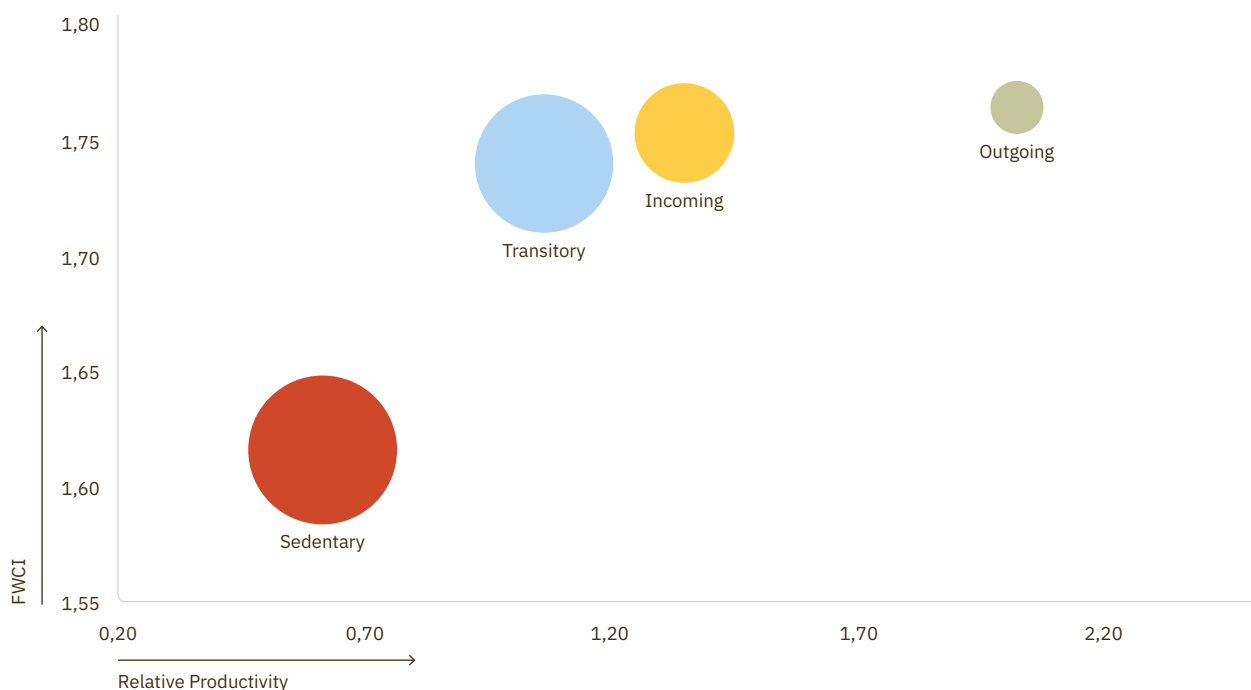


Figure 26. Migration patterns of AI researchers in the Netherlands: share of researchers vs productivity and impact. Outgoing researchers are the most productive and score highest on Field-Weighted Citation Impact. Source: Courtesy of Elsevier, AIC4NL

## Startups

The distribution of startups across the four layers of the AI stack is highly uneven.

AI stack layer	NL share of start-ups	EU average share of startups	NL share of capital invested	EU average share of capital invested	Key characteristics
Application-Value	89%	88%	76%	74%	Fast iteration, low CAPEX, UX-driven differentiation, high churn, sector expertise, regulatory trust, medium R&D intensity
Integration-Interface	8%	9%	4%	15%	Platform dynamics, community-driven adoption, long-term monetization
Model-Data	1%	1%	1%	6%	Winner-takes-most dynamics, massive R&D and compute spend
Hardware-Energy	2%	1%	18%	5%	Extreme capital intensity, hardware IP moat, long payback horizon

Figure 27. Distribution of startups and capital invested over the AI technology stack, for the Netherlands and Europe. Source: Dealroom, see Annex II

The table shows that both the Netherlands and the EU concentrate the vast majority of AI startups and capital investment in the application–value layer, where barriers to entry are low and time-to-market is fast. The hardware–energy layer stands out: while although only 2% of NL startups operate here, the NL allocates a disproportionately high 18% of total AI capital to this segment, far above the EU average of 5%. This indicates targeted Dutch investment in AI-enabling hardware despite a small number of firms. Foundational model builders and hardware builders are generally capital-intensive and dominated by a handful of global players.

The distribution looks this way because of the following reasons: (1) capital requirements rise exponentially toward the lower layers. Building a foundational model or designing chips requires tens to hundreds of millions in upfront investment; (2) talent and compute access act as gating factors:

only a few teams can realistically compete in those segments; (3) market entry at the top of the stack is faster, cheaper, and more demand-driven, which attracts a larger number of early-stage founders; and (4) infrastructure and deep-tech layers attract fewer but more heavily funded players, often with strong corporate, sovereign, or institutional backing.

## Application Domains

Comparing perceived opportunities for AI across multiple sectors in the Netherlands, sectors such as healthcare, defense & security, science, biotech & pharma, manufacturing & robotics, agriculture & food, and energy & climate are viewed as having particularly strong AI potential by Dutch AI startups.

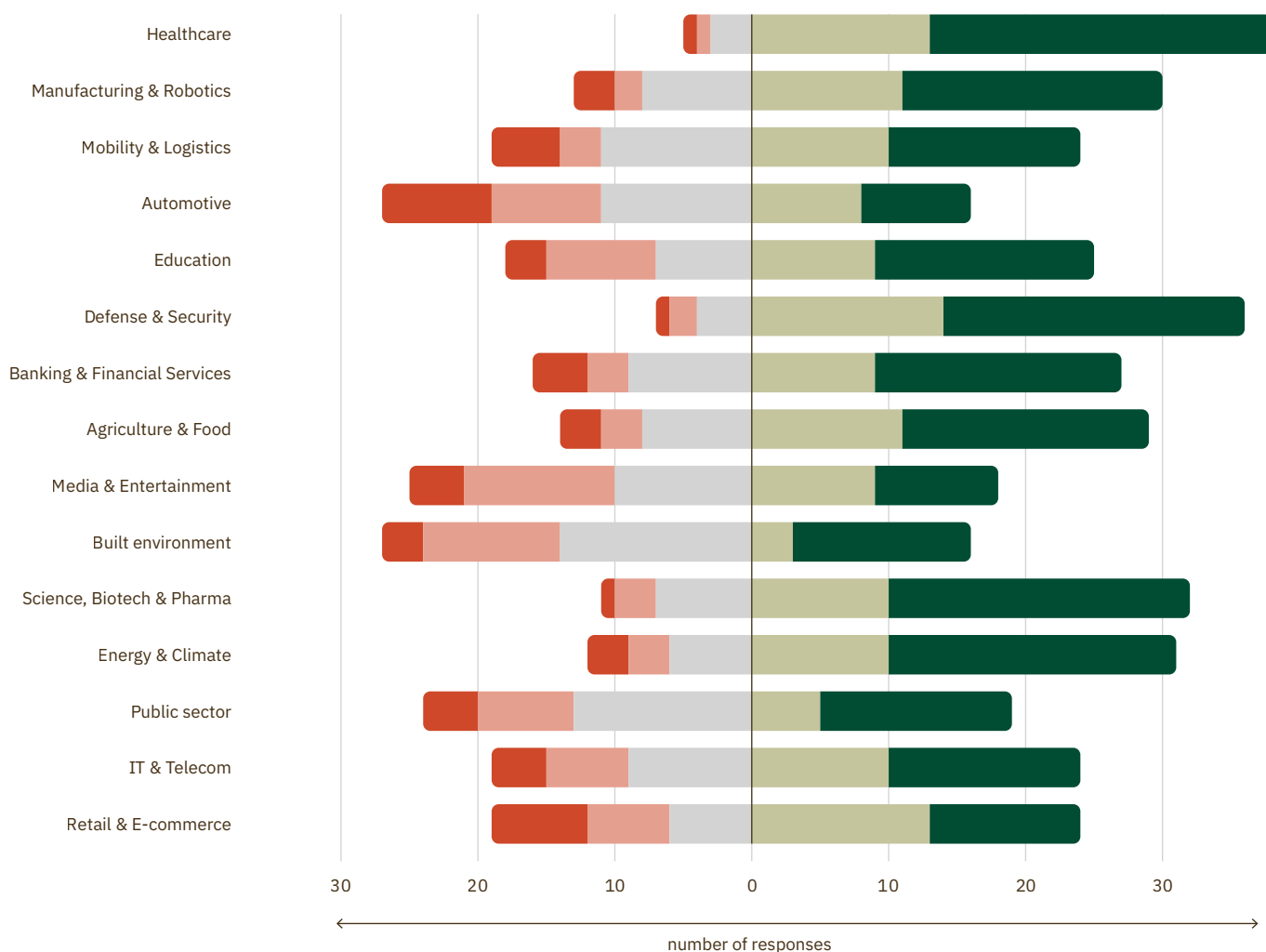


Figure 28. Where do Dutch AI startups see opportunities? Source: AI Deepdive Survey, October 2025

# Funding

Venture funding lags far behind that of the U.S. and China, and many startups face long lead times in securing private investments. According to interviewed experts, there exists a funding gap for Dutch AI startups to scale beyond seed investment. For example, the Amsterdam region ranks 6th in Europe for <\$15 million rounds, 7th for \$15-100 million rounds and 18th for \$100 million rounds for AI companies<sup>49</sup>.

This funding gap tends to lead Dutch founders to lower their ambitions for raising funds, which in turn widens the funding gap. Only 4% of Dutch AI founders surveyed plan to raise more than € 10 million<sup>50</sup>. This self-reinforcing loop is not productive for competition in a global AI market that operates at high execution speed and where fast scaling of compute and usage distribution are essential. Intervention can help closing this funding gap and raising fundraising ambitions simultaneously. Early-stage fund-of-fund matching would increase the number of Dutch AI startups and broaden the

pipeline for future series investments. Series A/B funds with tickets sizes of € 20-30 million should enable scale ups to cross the gap to profitability. To support new European champions that can globally compete in AI-hardware and foundation models, series B+ funding needs to be available with ticket sizes well over € 100 million.

# Defensible positions

A defensible position in AI is a feature that provides a company substantial advantage with respect to competitors. Four types of defensible positions can be discerned: access to data gathering and model validation through interfaces and usage; access to compute and energy; fast execution and short lead times by strongly funded teams to capture market first; and legal IP protections. According to Dutch AI startups, access to data and size of investments are most crucial. The first two types are unique to AI companies and drivers for a winners-take-most dynamic for scale, where good ideas attract capital,

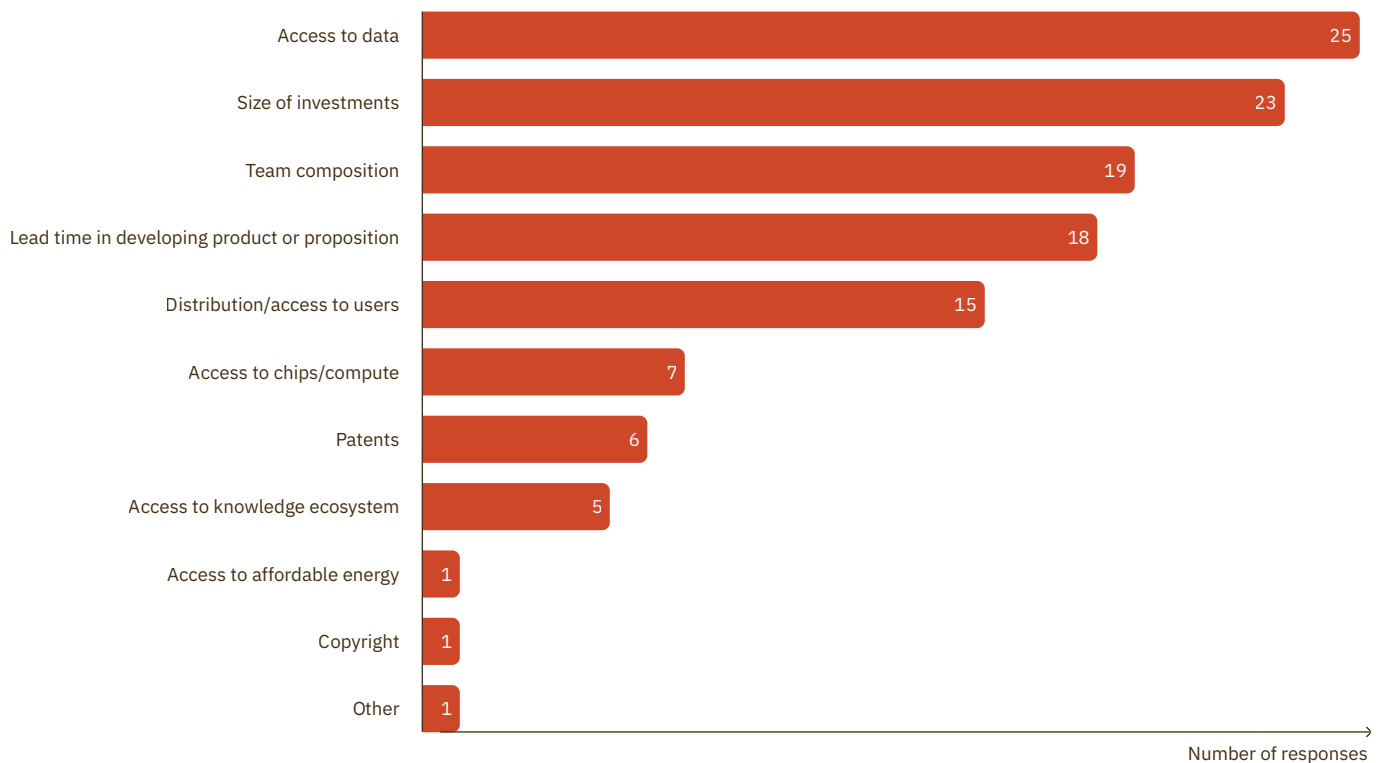


Figure 29. Defensible positions in AI, according to Dutch AI startups. Source: AI Deepdive Survey, October 2025

49. <https://www.prosus.com/our-insights/investment-insights/2025/the-netherlands-europes-hidden-ai-powerhouse-at-a-crossroads>  
50. <https://techleap.nl/reports/ai-scaling-challenges-for-dutch-founders-report>

capital enables access to resources, resources create powerful models, and powerful models attract more data, reinforcing and accelerating model improvement. This makes companies that successfully compete in the Model-Data layer unlike traditional software company growth and explains their strategic or sovereign value.

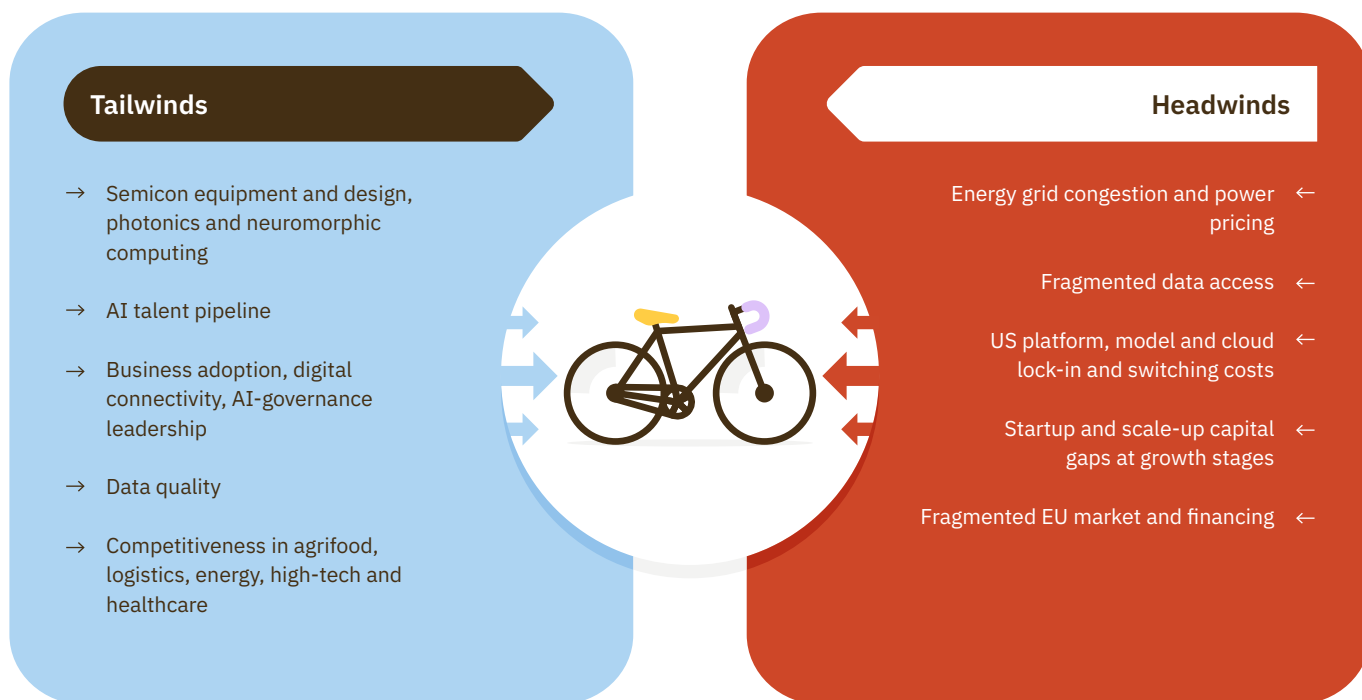


Figure 30. Headwinds and tailwinds for the Netherlands

# 7. Opportunities

## Vision

**AI rapidly transforms economies, across sectors, services and manufacturing.** Information work, a cornerstone in a service-oriented economy like the Netherlands, is progressively being automated, similar to how mechanical work was automated in the Industrial Revolution. The Dutch economy is at the cusp of major changes in the way work is organized. National security rests on being at the forefront of developing new technologies. If we fail to manage the rapid uptake of AI properly, the consequences could be severe.

**With AI as a new production factor, manufacturing can be automated in dark factories, routine work can be delegated to agents, and human attention can be directed to the most meaningful activities.** Where administrative burden is automated, doctors, nurses, police officers and teacher can devote time to what they consider core to their work, offering ways to tackle future shortages. Growth in labor productivity can make part-time work combined with caretaking for children and elderly within reach for anyone.

**Personalized AI assistants are quiet layer of intelligence woven into our lives.** Sensors and intelligence devices continuously monitor health biomarkers, detecting diseases years before symptoms appear. Cities dynamically adapt to human needs, optimizing traffic, energy, safety, and air quality.

**Generative AI unlocks revolutionary new medicines, sustainable materials, and advanced technologies.** AI is an indispensable tool for managing complexity, including complexity created with AI itself: from coordinating agents and swarms of drones, to monitoring model versioning or legal compliance, to generating more energy-efficient

algorithms and chip designs. AI enables us to use the gains of the digital revolution to clean up the mess of the industrial revolution<sup>51</sup>.

## Investment opportunities

**Trying to close the innovation gap with the U.S. or China in areas where they are winning today is a losing strategy.** Where there is uneven access to data, markets, chips, energy and capital, companies in the Netherlands will not be able to win a global race for scale through competition alone. Real opportunities lie not in what exists today but in building on strengths, anticipating likely technological directions, taking calculated risks, and investing in radical innovation. This needs to be complemented by an AI industrial policy that creates conditions for private investments and global competitiveness, with public data and compute resources anchored in long-term vision.

---

51. <https://www.ewmagazine.nl/economie/achtergrond/2025/04/ai-europa-economie-lezing-max-welling-1473456/>

Opportunity	Expected adoption in 2030	Market: economic opportunity	Market: lack of conventional financing	Strategic impact: reducing AI energy cost	Strategic impact: securing AI supply chain	Strategic impact: supporting societal transitions
Vertical applications	●●●●	●●●●	●●	●	●●	●●●●
Human and real-world interfaces	●●●	●●●	●●●	●	●●●●	●●
Data sharing and model validation platforms	●●●	●●●	●●	●●	●●●	●●●
Scientific AI	●●●	●●●●	●●	●●	●●	●●●●
AI accelerators, photonics and neuromorphic hardware	●●	●●●	●●●	●●●●	●●●●	●

● = A greater number of dots indicates a higher scale

Figure 31. Investment opportunities in AI

The categories in this table reflect a dual focus on market development and impact-driven investment. They capture market readiness and investment gap, to help assess when technologies are likely to mature, where significant economic potential exists, and where private capital may be hesitant due to high risk or long payback times. Strategic impact is considered where technologies enhance energy and resource efficiency, secure positions in the AI supply chain, or help solve major societal transitions such as in energy, healthcare, housing, agriculture or mobility.

**Vertical applications: companies that own proprietary (post-trained) models and data pipelines can create strong defensible positions.** This is essential in markets where competitive advantages are rapidly commoditized. Capturing market early helps securing usage distribution,

data accumulation and model validation. Non-vertical applications or model wrappers should have a different edge or defensible position to be considered a strategic investment. Strategic application opportunities exist where high-quality data exists in large volumes but is rare or difficult to access, and in application domains with strong socio-technical foundations such as data infrastructure, skilled talent, and rapid technology adoption. In the Netherlands these domains include agrifood, with innovative sensor technology, precision farming and high-tech greenhouse systems; logistics, with its strategic location in Europe, Rotterdam harbor, Schiphol airport and expertise in supply chain management; energy, with its offshore wind sector and smart energy grids; high-tech manufacturing, with its advanced engineering expertise, precision technology capabilities and innovative companies like ASML,

Philips and NXP; and healthcare, with its strengths in medical imaging and diagnostics, personalized medicine, and drug discovery.

**Human and real-world interfaces: the growth of data from intelligence devices will outpace the speed at which humans generate text.** Leveraging Dutch leadership in responsible AI, companies that develop AI's future interfaces secure strategic positions in data pipelines. The future of work will likely involve AI assisting humans in many tasks instead of taking them over completely, reinforcing the continued need for human understanding and control. AI will perceive and interact through multiple senses and data modalities beyond text, and will recognize human speech, emotions, and environmental cues. Opportunities are found in new interfaces, with neuro-symbolic AI, explainable AI, robotics or neural interfaces that bridge differences in how humans and AI process information in a human-centered way.

**Data sharing and model validation platforms: companies building data sharing and model validation platforms can create markets for dynamically capturing value.** Companies that offer data, model and reward environment exchanges create defensible positions. Structural financial incentives aligned with European legislation, could motivate businesses to share data and help reduce fragmentation of data accessibility. Where real-time interaction with the world or with humans is too slow and cumbersome, simulated reward environments are shaping up to be a key driver for training or post-training models with reinforcement learning. These proxies for human input or real-world interaction are poised to evolve into more information-rich and open-ended environments and rewards.

**Scientific AI: discovery in material, physical, geo, and life sciences - biology, pharma, and neural data - is shifting toward model-driven, simulation-rich workflows.** Significant opportunities are to be found where collaborations can be forged with leading Dutch academic research groups and scientific data sets. Through sensors, video, automated wet labs and other ways of interacting with the world, patterns found in nature can be efficiently modeled with machine learning. AI-accelerated discovery is a powerful tool for science by uncovering patterns in data that were previously

hidden and represents a paradigm shift in how research is conducted.

**AI-accelerators, photonics and neuromorphic hardware: though GPUs will likely remain a core tenet in AI for the near future, a hardware breakthrough could yield substantial energy-efficiencies and upend the GPU-dominated AI chip market.** Building on its strong existing semiconductor ecosystem, the Netherlands could emerge as a global AI leader in energy-efficiency, local computation and analog chip capabilities. More efficient AI decreases energy costs per operation, decreasing relative AI carbon footprint and lowering the barrier of market entry for AI startups. The Dutch ecosystem would be strengthened further with AI-chip production and moving up from Hardware-Energy to the Model-Data layer of the AI technology stack. This move could be driven by optimizing cross-cutting efficiencies in hardware-model alignment, for example specializing in edge hardware and models. Beyond energy-efficiency for general use cases such as tensor computation, inference or local data processing, a breakthrough could come from new hardware expressivity enabled by spike-dependent plasticity, thermal interaction, wave-based oscillations or other non-digital forms of computing that usher in a new era of AI beyond the statistical machine learning paradigm.

## Government

**Countries that succeed in aligning their innovation ecosystems, capital markets, regulatory frameworks, and public institutions are best positioned to benefit from AI-driven transformation.** AI has rapidly evolved into a strategic geopolitical asset. Nations capable of developing, scaling, and governing advanced AI systems increasingly shape global power dynamics; economically, militarily, and normatively. As a result, AI industrial policy has become central to questions of economic prosperity, societal resilience, and the preservation of democratic values. To reduce fragmentation across capital markets<sup>52</sup>, usage markets, and data accessibility it is necessary to act at a European scale. Regulation alone will not shape AI, as the Brussel Effect<sup>53</sup> is tied to a geoeconomic force that is losing global relevance.

52. [https://commission.europa.eu/topics/competitiveness/draghi-report\\_en](https://commission.europa.eu/topics/competitiveness/draghi-report_en)

53. <https://academic.oup.com/book/36491>

**For Europe - and particularly for the Netherlands, with its mid-sized, open, and service-oriented economy - the challenge is acute.** Our capital markets are comparatively shallow, we lack domestic technology giants capable of investing hundreds of billions of euros annually in AI infrastructure and research, and we are late in articulating an industrial policy tailored to the new geopolitical and technological era. The combination of external pressure and internal fragmentation demands coordinated action and bold strategic choices. This demands an AI industrial policy that mobilizes private investments, coordinates public policy areas in a European context, builds and funds hardware-software aligned AI roadmaps to overcome infrastructure and innovation gaps, and makes difficult choices based on strategy and foresight.

**An AI industrial policy should be executed with sufficient funding and scale specific instruments.** Increased access to tickets in the range of €20–30 million is crucial for many Dutch AI startups to cross the gap to profitability, and will help boost founders' fundraising ambitions. Ticket sizes of €100+ million are needed to grow local champions capable of competing globally in hardware and foundation models. Instruments require fast mechanisms that embrace risk, attract ambitious companies and reward radical innovation. National subsidies could be conditional on keeping meaningful company activities and IP in the Netherlands for a specified time. Challenge-based financing<sup>54</sup> allows public and private organizations to fund breakthrough and become launching customers for solutions.

**This needs to be backed by a strong public-private investment agenda, creating scale and removing barriers for pension funds, insurers and banks to participate in long-term, high-risk AI investments.** Europe is not lacking capital. In 2022, EU household savings were €1,390 billion compared to €840 billion in the US<sup>55</sup>. In the Netherlands alone, pension funds and insurers hold €2,400 billion in their portfolios. Through fiscal stimuli and public guarantees, the Netherlands could offer derisking incentives to institutional investors to invest in innovative AI ventures and public infrastructure. With Wet Toekomst Pensioenen, pension funds have greater flexibility to allocate capital toward illiquid

assets, allowing them to invest directly in AI, or indirectly via dedicated funds. This may enhance returns and diversify portfolios while building the country's future earning capacity.

**As a flanking policy for strategic investments, it is essential that the government removes structural barriers.** These including regulatory bottlenecks, infrastructure constraints such as grid congestion, comparatively high energy prices, and limited space for data centers. AI startups need competitive stock-option and tax regimes to attract top-tier talent. Sandboxes for responsible experimentation, where AI can be developed, tested, and scaled are crucial to build AI-products in a regulated environment.

**The Netherlands could develop a large-scale public AI deployment facility, but only if it forms a supportive pillar of its AI industrial policy.** The Netherlands could establish a Dutch AI Gigafactory<sup>56</sup> if it is feasible to train large public models that that can compete with private models, or where there is sufficient demand for specialized, smaller models or inferencing. This should be informed by scaling laws and costs forecasts for model training and inferencing<sup>57</sup>. Concerns about US export controls on GPUs or preferential access threatening national AI capacity could also be mitigated by joining forces with other European countries that investing in AI Gigafactories. Given the Netherlands' strong hardware design ecosystem, strong digital connectivity, presence of private AI-HPC providers, and currently unfavorable energy-grid and power-pricing environment, it could decide to build the AI Gigafactory of the future: a large-scale distributed deployment network optimized for edge-AI, energy-efficient hardware, and privacy-preserving local compute.

**The Netherlands should build public institutions that bring sensitive and vital data and digital infrastructure under democratic governance.** With AI becoming a cornerstone to the economy and citizens spending on average 6-8 hours per day interacting with digital technology, governments have a responsibility to protect sensitive data and vital infrastructure for its citizens and businesses. AI builds up personalized profiles from interactions with chatbots, personal assistant agents and

54. c.f. <https://www.aria.org.uk/>

55. [https://commission.europa.eu/topics/competitiveness/draghi-report\\_en](https://commission.europa.eu/topics/competitiveness/draghi-report_en)

56. [https://www.eurohpc-ju.europa.eu/european-commission-proposes-amendment-eurohpc-regulation-support-gigafactories-and-include-quantum-2025-07-16\\_en](https://www.eurohpc-ju.europa.eu/european-commission-proposes-amendment-eurohpc-regulation-support-gigafactories-and-include-quantum-2025-07-16_en)

57. <https://writing.antonleicht.me/p/datacenter-delusions>

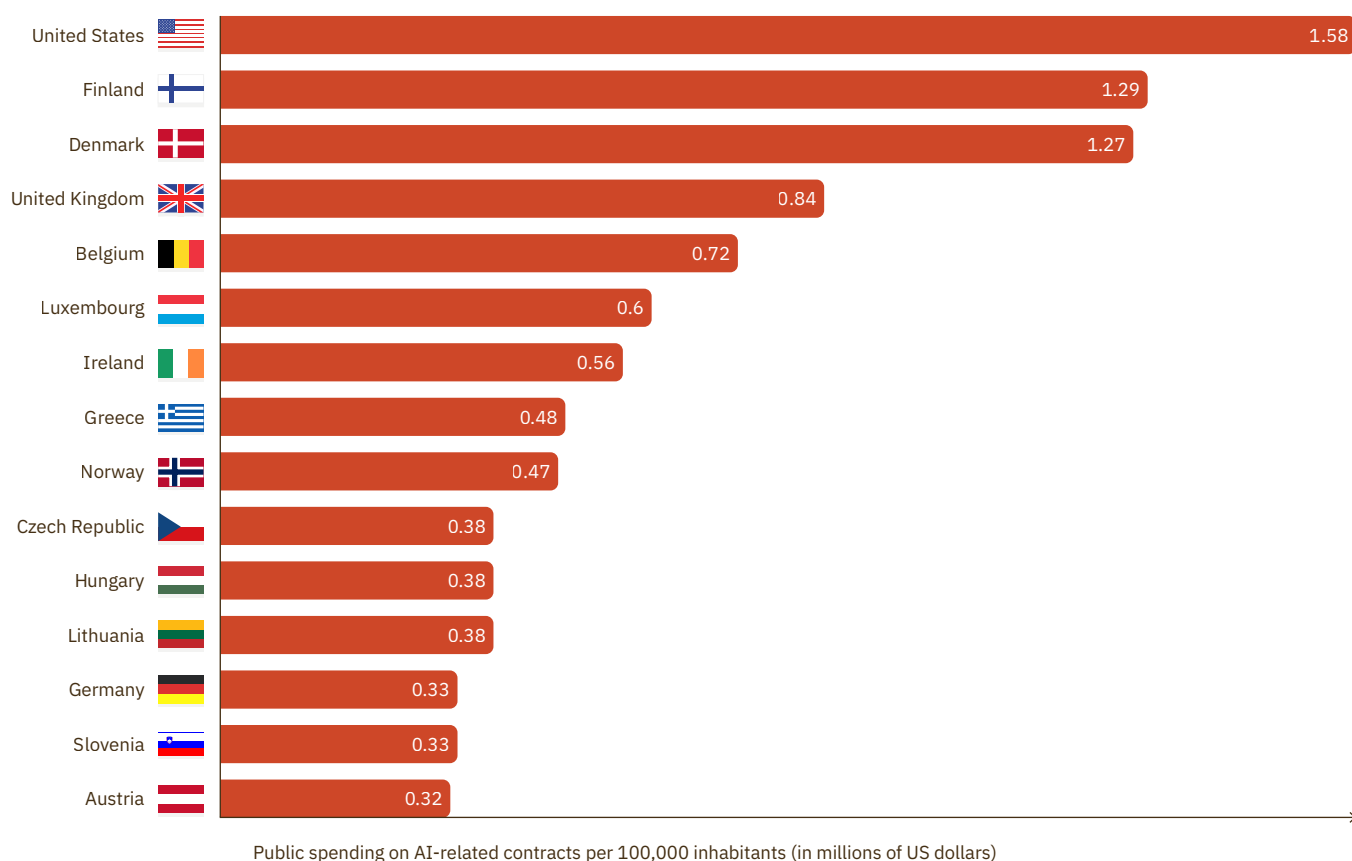
sensors, and will intimately get to know its users<sup>58</sup>. Sensitive data for key economic and defense activities are critical to national security. AI should fundamentally be a public-private technology. A Dutch public-private approach to AI should envision how AI can serve society, which data, models and infrastructure are crucial to keep under democratic governance, and by which entities these should be governed - national, European, public-private, commons, or steward-owned.

**An AI industrial policy should be reinforced by strong public demand and willingness to adopt AI.**

Public procurement represents a substantial investment potential and can be a driver of technological competitiveness and strategic autonomy. To reduce dependency on non-European providers and favor European AI solutions as a derisking strategy, procurement rules could embed increasingly stringent conditions for local

data storage, data privacy, interoperability, and prioritizing resilience and security over short-term cost savings. Defense, healthcare, research, education and public administration can have a crucial enabling role in reducing investment risk by being a launching customer and long-term commitments, ensuring predictable demand and lowering the perceived risks of early-stage innovation. Invest-NL could take a coordinating role in connecting startups with government demand, and work with European counterparts like Bpifrance and KfW to build a unified AI investment market.

Only through a coordinated, long-horizon approach can the Netherlands and Europe build AI capacity that is globally competitive, aligned with democratic values, and capable of supporting long-term prosperity and security.



**Figure 32. Public spending on AI-related contracts per 100,000 inhabitants in select countries, 2013-23 (sum).** Source: HAI 2025 AI Index Report

58. c.f. Yuval Noah Harari, Tristan Harris, “AI will know us better than we know ourselves” (<https://www.wired.com/story/artificial-intelligence-yuval-noah-harari-tristan-harris/>)

## Scenarios

**There are many uncertainties in how AI will develop in the coming years.** These include the speed at which capabilities grow through scaling foundation models, and how power over AI will be distributed<sup>59</sup>. These uncertainties can be monitored (see Annex I), combined with foresight and trends, and acted upon.

**Non-disruptive scenarios are unlikely.**

**If AI development slows and expectations of rapid automation fade as models hit limits in data, energy or scaling laws, there will still be major uptake and transitions enabled by current AI technology.** The Dutch government may continue to strengthen digital infrastructure and workforce skills, while supporting open science to prevent stagnation. Overregulating from the European AI Act may be avoided, either through a watered-down AI Act 2.0, delayed implementation or limited oversight and enforcement from national and European supervisory authorities. Europe realistically needs at least 2-3 competitive LLMs to keep value delivered by these models inside Europe.

**If model bottlenecks are overcome, highly capable AI systems will drive unprecedented productivity gains while likely centralizing power and economic influence.** Dominant non-European actors shape global markets, standards, and even elements of national security. The United States has already announced it intends to export its full AI technology stack - hardware, models, software, application, and standards - and expects adoption from its allies and partners<sup>60</sup>. The risk of dependency, inequality, and regulatory capture intensifies as the effects of widespread AI deployment rip through labor markets<sup>61</sup>. It is essential for the Netherlands to find points of leverage, for example by providing valuable data, model validation opportunities, manufacturing contributions, or owning elements of the AI hardware and foundation model supply chain.

**If Dutch government and industry jointly pursue a long-term strategy they can transform AI from a dependency risk into a catalyst for industrial renewal.** To mitigate short-term dependencies,

the Netherlands builds essential infrastructure for foundation models to retain knowledge, talent, and strategic autonomy together with partners in Europe. This is complemented by a robust AI industrial policy. The Netherlands spearheads a public-private approach to AI, owns vital elements of the global AI supply chain, and develops the next generation of AI that drives industrial and societal transitions while upholding core European principles of transparency, privacy, and accountability.

**The investment opportunities and government actions outlined in this chapter are relevant across these scenarios.** Opportunities are the strategic control points of the future AI technology stack: novel hardware, scientific AI, advanced data-gathering and validation interfaces and platforms, and vertical AI applications.

---

<sup>59</sup>. CFG report

<sup>60</sup>. full-stack export commitment from the US AI Action Plan

<sup>61</sup>. Anton leicht

# Afterword

This report reflects the many insights gathered through the thoughtful exchange of ideas. I thank the following experts for contributing their time and insights during interviews leading up to this report:

Matthieu van Amerongen	Nebul
Tijmen Blankevoort	Meta
Arjan van der Born	ROM NL
Bas Dunnebier	AIVD
Alexandre Ferreira Gomes	Clingendael
Rogier Fischer	Hadrian
Tamara Franssen	Techleap
Marcel van Gerven	Donders Institute
Bram van Ginneken	Radboudumc
Ruud Hendriks	Startupbootcamp
Anke Huiskes	NP-Hard Ventures
Corne Jansen	Inkef
Wim Kees Janssen	Syntho
Willem Jonker	AIC4NL
Arnold Juffer	Nebul
Daan Juijn	Centre for Future Generations
Vincent Kamphorst	Innovation Industries
Maurits Kaptein	TUe
Salar al Khafaji	Monumental
Douwe Kiela	Contextual AI
Durk Kingma	Anthropic
Michaël Kolk	Arthur D. Little
Bob van Luijt	Weaviate
Andy Lurling	LUMO Labs
Ard-Pieter de Man	Digital Holland
Archie Muirhead	IQ Capital
Maaïke Okano-Heijmans	Clingendael
Jelle Prins	Cradle
Bastiaan van de Rakt	AIXTERRA
Sweitze Roffel	IJCAI
Jörgen Sandig	Fermioniq
Cees Snoek	UvA
Maarten Stolk	Deeploy
Niels Taatgen	RUG
Gert-Jan Vaessen	Invest-NL
Anne Fleur van Veenstra	TNO
Bram-Ernst Verhoef	Axelera
Tom Wehmeier	Atomico
Sandra van der Weide	Ministerie van Economische Zaken
Max Welling	CuspAI
Peter Westerhuijs	TTT.AI
Christian van der Woude	Techleap
Jakub Zavrel	ZetaAlpha
Jelle Zuidema	UvA

I also thank the following persons who provided valuable help in shaping this report:

Bart van Campenhout, Sjoerd Dikkerboom, Liz Duijves, Chantal de Graaf, Esther Hogenhout, Eva Janssen, Fabian Kok, Wander van der Kolk, Elise Lie, Anouk Möller, Johan Stins, Mustafa Torun, Marloes Treffers and Ruben Wassink.

Stefan Leijnen

# Annex I: Investment hypotheses

Given the rapid pace of AI developments, it is recommended that Invest-NL continues to monitor global AI markets, technological trends and entrepreneurial opportunities with an annual or bi-annual update of this report. Prospects of paradigm shifts in machine learning will have direct implications for investment strategies. Early identification and support of promising ventures will be critical to secure a competitive position. Below a set of investment hypotheses that provide guiding assumptions for decision-making and a framework for identifying emerging trends.

**Data-first hypothesis:** the potential quality of a model is determined by the data used; how fast you approach this potential is determined by the available compute, in terms of energy, hardware and software.

**Bottleneck hypothesis:** the cost of AI will in the limit go down to the cost of energy for open data; the cost of AI will go down to the cost of energy and data in the limit for proprietary and commercial data.

**Universal-tensor hypothesis:** tensors are a fundamental representation of meaning; therefore GPU-type operations such as tensor multiplication and convolution will remain central to machine learning, and new processor types such as analog, neuromorphic, photonics or quantum computing are likely to use tensor-based representations of meaning.

**Universal-language hypothesis:** human language is a fundamental representation of meaningful reasoning and thinking; therefore models trained on non-language data, including world models, will continue to need language data.

**Monolithic-model hypothesis:** monolithic frontier models will eat into the application-value layer and smaller niche models, through extended reasoning, task and token length.

**Open-source hypothesis:** open-source models will continue to lag 6 to 12 months behind frontier proprietary models.

**Supply-chain hypothesis:** models will be considered strategic geopolitical assets like hardware, where nations consider control over a model's distribution key to economic and national security interests.

# Annex II: Notes on methodology

The AI Deep Dive survey was conducted between September 10th and October 22th 2025, supported by Invest-NL, ROM Nederland and Techleap. A total of 43 Dutch startups participated and completed the questionnaire: 14 startups indicated they are in Pre-seed phase, 21 Seed, 5 Series A/B, 2 later than Series A/B, and 1 Don't know/No answer. Out of the participants, 34 consider themselves active in the application-value layer, 28 in the integration-interface layer, 28 in the model-data layer, and 5 in the hardware-energy layer (multiple answers were possible). For the questions about model providers and funding gaps, participants could choose multiple answers. For the question about defensible positions, participants could choose a maximum of three features. We thank the survey participants for their contributions to this report.

The distribution of startups and capital invested over the AI technology stack is based on 10103 European AI companies gathered from Dealroom in August 2025. These entries were categorized into AI applications, interfaces and model and data hubs and new paradigm, foundation model, and infrastructure companies based on an LLM pipeline with category-specific prompts on shortened company descriptions gathering by web scraping. The data gathered through this experimental method is indicative, not conclusive, and is used here as a rough estimate to compare NL and EU companies and investments within the AI technology stack.

For the European market sizing, value is defined as producer surplus only discounting private capital WACC, base case shown at 10% nominal, with 8-12% sensitivities, portfolio mix for benefits is 40% for productivity/cost, 40% revenue growth, 20% risk/compliance; benefit timing: each year's spend yields a total benefit equal to BCR x spend, realized 20% in t+1, 50% in t+2, 30% in t+3; FX and inflation: euros only, nominal value. The EU InvestAI programme of €200 billion over 10 years is treated as already embedded in market expectations and no separate uplift is modeled.

The Artificial Analysis Intelligence Index for comparing Open Weights vs Proprietary models incorporates 10 evaluations: MMLU-Pro, GPQA Diamond, Humanity's Last Exam, LiveCodeBench, SciCode, AIME 2025, IFBench, AA-LCR, Terminal-Bench Hard, and t2-Bench Telecom.

# Annex III: Figure references

- Figure 1:** <https://hai.stanford.edu/ai-index/2025-ai-index-report>
- Figure 3:** [https://www.cs.rice.edu/~as143/COMP642\\_Spring22/Scribes/Lect6](https://www.cs.rice.edu/~as143/COMP642_Spring22/Scribes/Lect6)
- Figure 4:** <https://www.asimovinstitute.org/neural-network-zoo/>
- Figure 5:** <https://hai.stanford.edu/ai-index/2025-ai-index-report>
- Figure 7:** <https://hai.stanford.edu/ai-index/2025-ai-index-report>
- Figure 8:** <https://epoch.ai/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data>
- Figure 9:** <https://storage.googleapis.com/deepmind-media/Era-of-Experience%20/The%20Era%20of%20Experience%20Paper.pdf>
- Figure 10:** <https://iea.blob.core.windows.net/assets/601eaec9-ba91-4623-819b-4ded331ec9e8/EnergyandAI.pdf>
- Figure 13;** <https://www.iea.org/reports/energy-and-ai/>
- Figure 14:** <https://hai.stanford.edu/ai-index/2025-ai-index-report>
- Figure 15:** <https://artificialanalysis.ai/trends>
- Figure 16:** <https://evaluations.metr.org/gpt-5-report/>
- Figure 19:** <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai#/>
- Figure 20:** <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier>
- Figure 21:** <https://hai.stanford.edu/ai-index/2025-ai-index-report>
- Figure 23:** <https://hai.stanford.edu/ai-index/2025-ai-index-report>
- Figure 24:** <https://www.tortoisemedia.com/data/global-ai#rankings>
- Figure 32:** <https://hai.stanford.edu/ai-index/2025-ai-index-report>

# About the author



## Stefan Leijnen

Stefan Leijnen has worked for over two decades at the intersection of artificial intelligence research, innovation, and public-policy strategy. He has been involved in national and European initiatives that work to strengthen economic resilience and technological sovereignty with AI.

Son of a genetic algorithms expert at Philips, Leijnen began his career in academic AI research. He wrote his dissertation on creativity and emergence in machine learning. Over the years he has worked at the University of California, Berkeley, Tufts University and the University of British Columbia, contributing to fields such as machine learning, artificial creativity, and autonomous systems. In 2016 he founded the Asimov Institute, a non-profit research lab on creative AI.

Since 2019 he leads the international team at the AI Coalition for the Netherlands (AIC4NL) and is professor of the AI research group at Utrecht University of Applied Sciences. He has collaborated closely with research institutes, startups, and applied-science organizations to bridge the gap between fundamental AI knowledge and real-world deployment. He is motivated by a belief that the shape of technology shapes the future.

As an advocate for AI sovereignty and responsible innovation, Stefan contributes regularly to national and EU-level dialogues on AI strategy, investment, governance and industrial readiness. In this AI deep-dive he draws on his vantage point - from research lab to strategy table - to guide you through how AI is developing, where competitive advantages lie, and how strategic investment can create sustainable value.

# Copyright & disclaimer

This research report is provided for information purposes only. Although it has been prepared with the greatest possible care, any use of the information is at your own risk. Invest-NL accepts no responsibility or liability for any damage, loss, or inconvenience resulting from the use of this report or the data contained herein.

Portions of this publication may include graphs, data points, or analyses from external sources. These have been reproduced with appropriate source attribution. Invest-NL does not guarantee the completeness, accuracy, or continued availability of such external data.

All intellectual property rights related to this report are owned by Invest-NL, its licensors, or the sources referenced. If you wish to redistribute, reproduce, or otherwise reuse this publication (in whole or in part), please inform us in advance.

For more information, please contact Invest-NL via [info@invest-nl.nl](mailto:info@invest-nl.nl).

# INVESTNL

**Invest-NL**

Kingsfordweg 43-117

1043 GP Amsterdam

T +31(0)882036700

[www.invest-nl.nl](http://www.invest-nl.nl)

[info@invest-nl.nl](mailto:info@invest-nl.nl)

© 2025 | Invest-NL N.V.