FOUNDATIONS OF ARTIFICIAL INTELLIGENCE: BIG DATA, PROCESSORS AND LARGE LANGUAGE MODELS

Prof. Dr. Şeref Sağıroğlu
Gazi University

Attribution: Sağıroğlu, Ş. (2026). Foundations of artificial intelligence: Big data, processors and large language models. In H. Arslan, A. Taşdelen, E. Kuzucu Hıdır, & O. A. Çetin (Eds.), *Artificial intelligence and national technology reflections* (pp. 59–82). Turkish Academy of Sciences. https://doi.org/10.53478/TUBA.978-625-6110-86-1.ch03

10.53478/TUBA.978-625-6110-86-1.ch03

FOUNDATIONS OF ARTIFICIAL INTELLIGENCE: BIG DATA, PROCESSORS AND LARGE LANGUAGE MODELS

Prof. Dr. Şeref Sağıroğlu
Gazi University
President of AI Ecosystem Association of Türkiye

Abstract

This chapter reviews the interdependencies among AI-specialized microprocessors, big data ecosystems, and large-scale foundation models, including large language models (LLMs) and large multimodal models (LMMs). The primary objective is to analyze how the convergence of data availability, algorithmic advancements, and hardware innovation has reshaped the artificial intelligence (AI) landscape and accelerated the development of generative AI systems. Drawing on industry reports, technical documentation, and recent literature published between 2024 and 2025, this study examines three critical dimensions: (i) the evolution of processing architectures from general-purpose CPUs to specialized accelerators such as GPUs, TPUs, NPUs, and emerging paradigms (i.e., neuromorphic and photonic computing), (ii) the role of large-scale and multimodal big data in enabling foundation model training, and (iii) the growing energy consumption and sustainability challenges associated with large AI systems. In addition, national and global semiconductor strategies, including investments in exascale supercomputing and sovereign AI infrastructures, are analyzed to understand geopolitical and technological dynamics. The findings highlight three key insights. First, the co-evolution of big data, processors, and model architectures forms the foundational driver of modern AI capabilities. Second, the rapid scaling of LLMs and LMMs has significantly increased computational and energy demands, making hardware efficiency and infrastructure design critical bottlenecks. Third, global competition in semiconductor technologies and AI infrastructure is intensifying, with nations prioritizing technological sovereignty and secure access to advanced computing resources. In conclusion, the future of AI depends on a balanced integration of scalable data ecosystems, energy-efficient hardware, and advanced model architectures. Sustainable and secure AI development will require coordinated efforts in hardware–software co-design, algorithmic optimization, renewable energy integration, and policy-driven governance to ensure responsible and inclusive technological progress. Modern AI relies on three key interconnected pillars: LLMs for advanced language processing, LMMs for handling diverse data types (i.e., text, images, audio, video) beyond language alone, and AI-specialized microprocessors (i.e., GPUs, TPUs, custom accelerators) that supply the massive computational power needed for training and running these models efficiently. Together, these enable foundation models capable of broad, generalized learning and real-world applications while posing significant challenges in energy use, scalability, and deployment as observed in research and industry trends from 2024–2025.

Keywords

Large Language Model, Hardware, Microprocessor, Big Data, Generative AI

Introduction

Artificial Intelligence (AI) has a history spanning more than 75 years, beginning with early theoretical work in computational logic and neural networks in the 1950s and evolving into the powerful, data-driven systems of today. Early AI research of the *perceptron* in the 1950s laid the groundwork for learning from data and performing simple classification tasks, but it wasn't until advances in algorithms, data availability, and computing hardware that performance began scaling to practical levels.

The transformative success of modern AI arises from the synergy between massive datasets, advanced algorithms (especially deep learning -DL- architectures such as transformers), sophisticated processing hardware, and sustained R&D investment. In particular, computational power has always been a fundamental requirement across the full lifecycle of AI workflows: training, development, analysis, testing, and deployment. Whereas early AI systems could run on general-purpose CPUs, contemporary AI models, especially Large Language Models (LLMs) and recently Large Multi-Modal Models (LMMs/LMMMs), depend on highly parallel computing architectures to process vast amounts of data in different formats efficiently. GPUs (Graphics Processing Units) evolved from graphics-oriented chips to AI workhorses capable of performing thousands of simultaneous operations necessary for tensor algebra, which is at the heart of DL training and inference, outperforming CPUs by orders of magnitude in parallel workloads.

To further accelerate neural computations, industry pioneers such as Google introduced Tensor Processing Units (TPUs), application-specific integrated circuits (ASICs) specifically optimized for the matrix multiplications and tensor operations that dominate neural network training and inference (Google, 2017). Modern foundation models like GPT, Gemini, Claude, Falcon, LLaMA, and similar systems are trained on big datasets encompassing trillions of tokens comprising text, code, and multimodal data and requiring enormous compute resources. For example, training of large transformer-based models involves thousands of GPUs working in distributed clusters over weeks or months, consuming extensive energy and computational cycles to adjust billions to trillions of parameters and to learn patterns from world-scale datasets (HAI, 2025). The availability of such large datasets and the development of high-performance semiconductor technologies have thus been central drivers of AI progress, enabling models that can learn, understand and generate human-level language, vision, and other modalities, as well as power diverse applications across science, industry, and everyday digital services (Sparkes, 2025).

Type and generation of GPUs used in AI systems certainly play a critical role in determining the feasibility, efficiency, and scalability of training large models, particularly LLMs and LMMs. The transition from CPUs (scalar operations) to GPUs

(vector operations) significantly facilitates the processing and analysis of massive datasets. In addition, designing specialized chips like TPUs, NPUs (Neural Processing Units) and DPU (Data Processing Units) today help us to analyze big data or to train large neural models with the help of huge or big data collected from all over the world. More advanced processing units (PUs) are always expected for scientific, sectoral or industrial of AI developments using available processing units or new extended version or new developments such as GPU, TPU, BPU (Brain Processing Unit), QPU (Quantum Processing Units), NPU, DPU, Analog Resistance RAM (Analog RRAM), etc. (NVIDIA, 2022; 2023; 2024a; 2024b; 2024c; 2024d; Google, 2017; Microsoft 2023a; Zuo&Sun, 2025; AMD 2024a, 2024b, 2024c; TSMC 2024a, 2024b) as shown in Figure 1.

Developing vast amounts of PUs provides not only to train and running of large or complex AI models but also to serve real-time or near real-time applications and implementations. Today's major and influential modern neural networks including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Generative Pre-trained Transformers (GPT) (as shown in Figure 2) mainly require the given processing units. Training these sequence models (RNN, LSTM) and generative models (GAN, VAE, GPT) models requires substantial computational resources and large-scale processing units to handle massive datasets, enabling the extraction of meaningful representations from complex data structures and advanced data analytics.

Figure 1

Mostly Available and Underdeveloped Processors in the Market and Literature

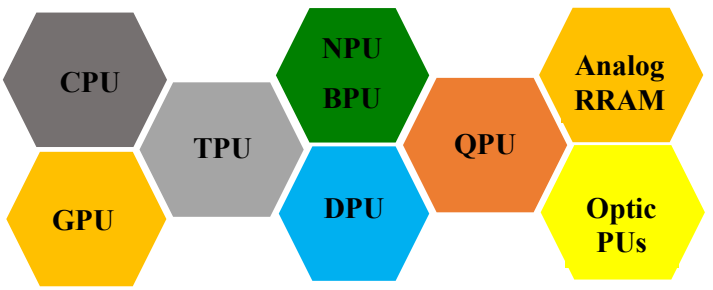
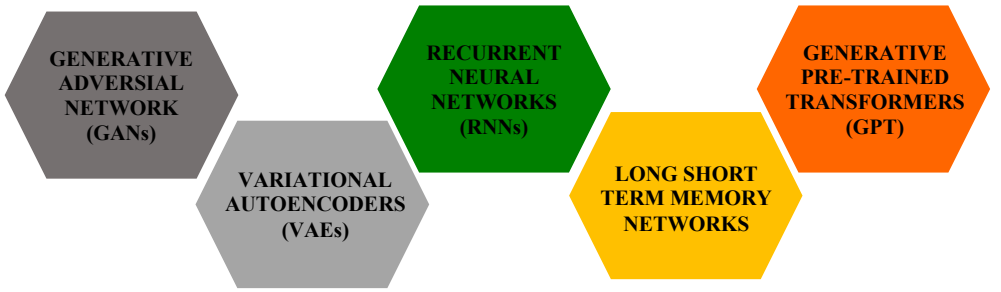


Figure 2

Recently Used and Classified Sequential and Generative AI Models



The emergence of specialized hardware, including TPUs, NPUs, and DPUs, has further optimized neural network operations, allowing efficient handling of massive, multimodal datasets. Training state-of-the-art models now requires distributed computing infrastructures with thousands of processing units operating over extended periods, consuming substantial computational and energy resources. Advances in semiconductor technologies and the continuous development of new processor architectures (i.e., QPUs, BPUs, and analog memory systems) are driving the scalability and performance of AI systems. These technologies not only enable the training of complex sequence models and generative models but also support real-time and large-scale AI applications across diverse domains. In this context, the co-evolution of microprocessor technologies, big data ecosystems, and advanced AI models is shaping a new computational paradigm that will define the future capabilities, scalability, sustainability and societal impact of intelligent systems.

When AI strategies and action plans are reviewed, the common sentences are mostly about AI models (LLMs/LMMs), processors, infrastructure (data center), and also competition, leadership, building an ecosystem, influence over global standard, economic growth, privacy and security issues. The integration of LLMs, advanced processors, and big data ecosystems is redefining national power in the digital age. Countries that invest strategically in these domains will not only achieve technological leadership but also secure economic resilience, national security, and long-term sovereignty in an increasingly AI-driven world. This integration and its ecosystem are briefly discussed in the context of Türkiye. This chapter consists of 5 Sections. Section 2 introduces the importance of the combinations of large language models, processors and big data world. Section 3 focuses on hardware and processors. Section 4 explains data centers and super computers. Energy demand and consumption of AI and Generative AI (GenAI) systems are summarized in Section 5. Increasing hardware requirements and mitigation strategies are explained in Section 5. Finally, the work is concluded in Section 6.

1. Combinations of AI Ecosystem (Models, Processors and Big Data)

LLMs and LMMs are known as foundational models and mostly require huge computational powers that enable the transformation of heterogeneous data types such as text, images, audio, and video into high-dimensional vector representations and facilitate semantic relationships across these modalities (Xue et al., 2024; Chen et al., 2024). Recent advances in models have significantly increased the demand for both large-scale datasets and intensive computational resources, necessitating the development of novel training algorithms and optimization techniques.

The growing computational requirements of training machine learning (ML), deep learning (DL) and GenAI models on massive and diverse big data sources have fundamentally reshaped semiconductor architectures and manufacturing processes. In this context, the complexity among model architectures, processing hardware, data availability, and semiconductor innovation has emerged as a cornerstone of modern AI systems.

Even if the convergence of algorithmic innovation, designing advanced processors, large-scale data and data ecosystems, and cutting-edge semiconductor manufacturing represents one of the most profound technological transformations of AI age, today's AI ecosystem is extended to topics as shown in Figure 3 encompassing data resources, research and innovation activities, algorithms and models, computing power and infrastructure, application domains and objectives, human capital, as well as ethical and legal frameworks. The sustainability and effectiveness of future AI systems critically depend on the balanced co-evolution of these interdependent components.

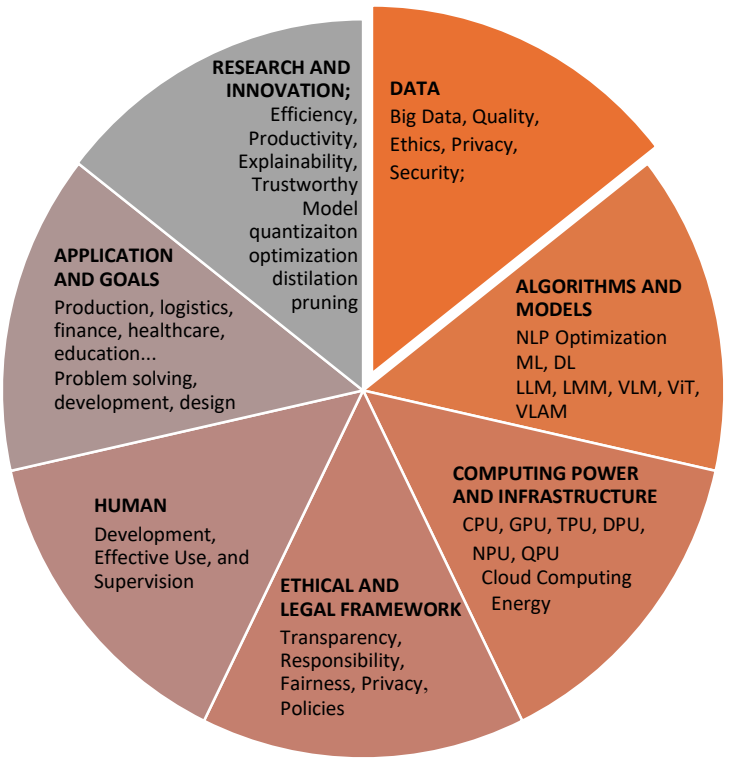
This ecosystem was notably accelerated by Google's seminal 2017 model of *"Attention Is All You Need"* proposed by Vaswani et al. (2017). The proposed transformer model relies exclusively on attention mechanisms and eliminating both recurrent and convolutional structures and achieves a BLEU score of 28.4 on the WMT 2014 English-to-German translation task, while significantly reducing training time compared to prior state-of-the-art models. This breakthrough laid the foundation for today's large-scale models.

Contemporary architectures including LLMs, LMMs, VLMs (Vision Language Models), and VLAMs (Vision Language Action Models) now comprise billions of parameters, with emerging systems approaching trillion-parameter scales. Training and deploying such models require massive parallel processing capabilities that exceed the performance limits of traditional CPU-based systems. Consequently, modern AI workloads increasingly rely on specialized accelerators such as GPUs, TPUs, NPUs, DPUs, and emerging paradigms including QPUs (See Figure 1).

The extreme computational intensity of modern foundation models has driven rapid innovation in specialized hardware accelerators and advanced semiconductor manufacturing technologies. As of 2024, GenAI has become a primary driver of semiconductor industry growth, with AI-focused chips contributing a mid-teens percentage of leading foundries' revenues and pushing advanced fabrication nodes to near or beyond full capacity utilization. Supporting this new AI-driven computing paradigm requires a vertically integrated technology stack ranging from nanoscale transistor physics to cloud-scale distributed computing infrastructures, each layer introducing distinct challenges and opportunities for innovation.

Figure 3

Combinations of Overall AI Ecosystem



2. Hardware and Processors

The rapid evolution of LLMs and LMMs is closely tied to breakthroughs in computational power, particularly in supercomputing systems, advanced hardware platforms, and chip design technologies. This section summarizes the current state of these developments in the world and studies achieved in Türkiye.

2.1. Hardware Platforms for Large Model Development

The rapid evolution of processing units has profoundly transformed the AI landscape, giving rise to unprecedented computational demands that are reshaping semiconductor design, architecture, implementation, and application domains (Top500, 2023; Global Growth Insights, 2025). This section examines relationships tightly among large model architectures, processing technologies, and semiconductor innovations covering algorithmic innovation, advanced processing units, and large-scale data availability.

The computational requirements of AI foundation models have directly driven the development of specialized processing units, hardware accelerators, and advanced semiconductor manufacturing processes, particularly within cloud-based computing environments. As of 2024 and 2025, AI has emerged as the dominant driver of semiconductor innovation, with reports indicating that AI-oriented chips are being manufactured at or beyond full capacity utilization at advanced technology nodes (Yao et al. 2024, Ren et al., 2024, Global Growth Insights, 2025). Supporting contemporary AI workloads necessitates a vertically integrated technology stack, spanning nanoscale transistor physics, chiplet-based architectures, high-bandwidth memory systems, and cloud-scale distributed computing infrastructures introducing new avenues for performance, optimization and innovation.

The hardware infrastructure required for the development and deployment of foundation models primarily consists of high-performance accelerators such as GPUs, TPUs, NPUs, and other domain-specific computing architectures. Given the rapid pace of innovation in this field, it is essential to closely monitor technology developments led by major semiconductor and system designers, including Intel, NVIDIA, Qualcomm, MediaTek, Samsung, and several emerging manufacturers from East Asia, particularly China and South Korea. Among currently available platforms, NVIDIA GPUs dominate large-scale LLM training and inference workloads. Widely adopted products include the V100, A100, H100, H200, and the recently introduced B200 accelerators (NVIDIA, 2017; 2020; 2022; 2023; 2024a; 2024b; 2024c). These architectures have progressively expanded computational throughput, memory bandwidth, and interconnect performance to meet the extreme demands of transformer-based models. These are briefly explained below:

- NVIDIA V100 (NVIDIA, 2017), based on the Volta architecture, introduced Tensor Core technology optimized for mixed-precision matrix operations. Available in 16 GB and 32 GB memory configurations, a single V100 GPU delivers computational performance comparable to dozens of traditional CPU cores, making it suitable for AI, high-performance computing and graphics workloads.

- NVIDIA A100 (NVIDIA, 2020), introduced in 2020 with the Ampere architecture, significantly advanced AI training efficiency by incorporating TensorFloat-32 (TF32) precision. Equipped with 432 Tensor Cores and 6,912 CUDA cores, the A100 provides up to six times higher AI training performance compared to Volta-based GPUs. With high-bandwidth memory configurations ranging from 40 GB to 80 GB, it has become a cornerstone platform for large-scale deep learning, natural language processing, and speech recognition tasks.
- NVIDIA H100 (NVIDIA, 2022), based on the Hopper architecture and released in 2022, represents a major leap in accelerator design. Featuring fourth-generation Tensor Cores, NVLink interconnect technology, and support for optimized LLM inference frameworks such as TensorRT-LLM, the H100 delivers approximately five to nine times higher performance than the A100 in transformer workloads. Its architecture integrates 18,432 CUDA cores, 640 Tensor Cores, and 80 Streaming Multiprocessors, enabling efficient GPU-to-GPU communication and large-scale parallelism essential for modern LLM training.
- NVIDIA H200 (NVIDIA, 2023), announced in late 2023, extends the Hopper family by significantly increasing memory capacity and bandwidth. With approximately 141 GB of high-bandwidth memory and multi-terabyte-per-second data transfer rates, the H200 is specifically optimized for generative AI and large-context LLM workloads that require processing massive parameter sets and datasets.
- NVIDIA B200 (NVIDIA, 2024a; 2024b; 2024c), based on the Blackwell architecture and introduced in 2024, represents a paradigm shift in AI accelerator design. Comprising approximately 208 billion transistors and capable of delivering up to 20 petaflops of AI performance, the B200, particularly within the GB200 Grace-Blackwell superchip configuration, is reported to achieve up to 30×performance improvements over the H100 while substantially reducing cost and energy consumption per operation. This architecture is designed explicitly to support trillion-parameter-scale foundation models.
- Beyond NVIDIA, other processor families are increasingly contributing to the AI hardware ecosystem. AMD's Ryzen AI 300 Series processors integrate NPUs delivering up to Top50, targeting on-device AI inference and next-generation personal computing workloads (AMD, 2024a). Similarly, AMD's Ryzen 9000 Series, based on the Zen 5 microarchitecture, offers high core counts and large cache capacities for hybrid AI and general-purpose workloads (AMD, 2024b).

- Custom architectures such as AMD's XDNA NPUs illustrate how AI capabilities are increasingly embedded in consumer and edge devices (AMD, 2024c).
- 2024–2025 developments emphasize custom silicon for AI workloads. For example, Google's Ironwood TPU targets inference acceleration (Google, 2024), while Microsoft's Maia 200 aims to compete in both training and inference computational efficiency (Microsoft, 2023).

Emerging AI chip developers are also gaining prominence. Rebellions, a South Korean startup founded in 2020, has introduced several AI accelerators, including the Ion, Atom, and Rebel chips, targeting financial services, data centers, and LLM-centric workloads (Rebellions, 2023). Likewise, DEEPX, another South Korean company established in 2018, focuses on energy-efficient on-device AI accelerators optimized for next-generation LLM inference across embedded and edge platforms (Rebellions, 2024; DEEPX, 2024).

At the manufacturing level, Taiwan Semiconductor Manufacturing Company (TSMC) remains the dominant global foundry, producing advanced AI chips for leading designers such as NVIDIA and Apple. TSMC's roadmap toward 2-nanometer process technologies is expected to further enhance performance, energy efficiency, and transistor density for future AI accelerators, including those intended for defense and critical infrastructure applications (TSMC, 2024a; 2024b).

Unlike electronic chips, photonic AI chips exploit optical signal propagation to perform computation and data transfer, enabling massive parallelism across multiple wavelengths while significantly reducing heat generation (Shen, et al., 2017; Miscuglio and Sorger, 2020). This paradigm supports high-throughput operations essential for AI workloads, including neural network inference, real-time pattern recognition, and image analysis. Recent breakthroughs demonstrate that computation using light instead of electrical signals can achieve up to two orders of magnitude higher energy efficiency compared to traditional semiconductor-based AI accelerators (Market Report Intellect, 2025). These advances position photonic AI chips as a promising solution to the growing sustainability challenges faced by large-scale AI systems and data centers and expected to rise to $5.6 billion by 2033 and increase at a CAGR of 22.5% from 2026-2033.

Neuromorphic processors are also another significant solution (Intel, 2021; Global Growth Insights, 2025; Open-neuromorphic, 2025). They are a class of brain-inspired computing architectures designed to emulate the structure and function of biological neural systems. Unlike conventional von Neumann architectures, neuromorphic systems use event-driven spiking neural networks (SNN) where computation and memory are co-located and only activated upon significant input spikes. This results in substantially lower power consumption, parallel processing, and real-time

adaptive behavior, making them particularly attractive for edge AI, robotics, autonomous systems, and sensor networks. Several neuromorphic processors and systems are at the forefront of research and early deployment such as Intel Loihi 2 implementing highly scalable SNNs with real-time on-chip learning and low energy per spike (Intel, 2021; Yao et.al. 2024); BrainChip Akida and Akida Pulsar of digital neuromorphic inference with orders-of-magnitude reductions in energy and latency for edge AI tasks, evolving into cloud-accessible frameworks (BrainChip, 2025); Innatera Pulsar & Mixed-Signal Systems integrate SNN engines with conventional microcontrollers to support ultra-low-power sensing and intelligent edge applications (Open-neuromorphic, 2025) and Large-Scale Systems such as *SpiNNaker 2* and brain-inspired supercomputers (Genkina, 2024). Collectively, these hardware platforms form the technological backbone of contemporary large model development, enabling scalable training, efficient inference, and the secure deployment of large-scale foundation models across cloud, edge, and on-device environments.

2.2. Hardware and Chip Development in Türkiye

In recent years, Türkiye has intensified its efforts to establish a sustainable and strategically independent semiconductor and chip development ecosystem. These initiatives are primarily driven by national priorities in defense, healthcare, AI, and critical infrastructure, where domestic chip design and fabrication capabilities are increasingly regarded as essential for technological sovereignty and supply chain resilience. One of the leading research infrastructures in this field is the Micro-Electro-Mechanical Systems Research and Application Center (METU MEMS) at Middle East Technical University was founded in 2008 (METU MEMS, 2026). The center was established on the foundation of the microelectronics cleanroom facility originally developed within the TESTAŞ project and subsequently transferred to METU in 1998. Since its inception, METU MEMS has developed into a nationally recognized hub for microfabrication and semiconductor research. To date, the center has secured over $70 million in research and development funding through projects supported by both public institutions and private-sector partners.

Complementing academic research efforts, Türkiye has also taken concrete steps toward establishing domestic chip manufacturing capacity. In 2023, investments were initiated to develop a Chip Production Facility at the TÜBİTAK Gebze Campus, marking a significant milestone in the country's semiconductor roadmap. In parallel, TÜYAR, a company founded under the auspices of the Presidency of Defense Industries (SSB), was established to coordinate and implement national chip production strategies. TÜYAR's mandate includes prioritizing Türkiye's semiconductor policies, fostering public–private partnerships, and facilitating strategic investments aimed at strengthening domestic chip design, fabrication, and packaging capabilities.

Finally, Aselsan and Tübitak-Bilgem have successfully developed “çakıl”, marking its debut as Türkiye’s first indigenously designed national processor. The chip is built using a 65nm manufacturing process, operates at a 400 MHz clock speed, and utilizes the open-source RISC-V instruction set architecture. It is specifically engineered to provide a secure, domestic computing heart for critical defense platforms such as UAVs and flight control units. The success of this initial single-core version has paved the way for current research into multi-core processors designed to accelerate AI tasks (Tübitak-Bilgem, 2026).

3. Data Centers and Super Computers

These three key insights are most concretely realized within AI data centers, which function as the central integration layer of modern AI ecosystems. First, the co-evolution of big data, processors, and model architectures is operationalized in data centers, where massive datasets are stored, managed, and processed using highly parallel accelerator clusters optimized for large-scale model training and inference. Second, the rapid scaling of LLMs and LMMs directly translates into exponentially increasing demands on data center infrastructure, including compute density, memory bandwidth, cooling systems, and energy supply. As a result, AI data centers have become critical bottlenecks where hardware efficiency, thermal management, and energy-aware scheduling determine the feasibility and cost of deploying large-scale AI systems. Third, the intensifying global competition in semiconductor technologies is increasingly reflected in the race to build sovereign AI data centers, enabling nations to secure independent access to high-performance computing resources, protect sensitive data, and support the development of national-scale LLMs and LMMs.

In addition, supercomputers represent the highest-performance realization of the same ecosystem, embodying the convergence of big data, advanced processors, and large-scale AI models at extreme scale. First, the co-evolution of data, hardware, and model architectures is most evident in supercomputing systems, where exascale infrastructures integrate massive datasets with tightly coupled GPU/accelerator clusters to enable the training of trillion-parameter LLMs and LMMs as well as large-scale scientific AI applications. Second, the rapid scaling of foundation models amplifies computational and energy demands to levels that only supercomputers can sustain, making them critical platforms for benchmarking hardware efficiency, parallelization strategies, and system-level optimization. These systems highlight the limits of current architectures in terms of power consumption, cooling, and interconnect performance. Third, the global competition in semiconductor technologies and AI infrastructure is strongly reflected in national investments in exascale and beyond-exascale supercomputers, which serve as strategic assets for achieving technological sovereignty, advancing scientific discovery, and developing secure, nation-scale AI models.

In this context, AI data centers and super computers are no longer passive infrastructure components but have evolved into **strategic assets**, shaping technological leadership, economic competitiveness, and the sustainability of the next generation of AI systems. The United States, Europe, Japan and China have made substantial investments in leadership-class supercomputers designed explicitly for AI and scientific workloads, as given below (Top500, 2023):

- **ORNL Frontier Supercomputer (ORNL, 2023):** Achieved 1,194 ExaFlops first then sustained performance improvements to approximately 1,353 ExaFlops. Primarily used for climate modeling, fusion research, astrophysics, and AI-accelerated science. Became the world's first operational exascale system, supporting large-scale scientific simulation and AI workloads in the USA.
- **Discovery Supercomputer (DOE, 2024):** Under development by the U.S. Department of Energy, projected to be three to five times faster than Frontier, targeting AI-native scientific computing, digital twins, and large multimodal models for science.
- **Aurora Supercomputer (ANL, 2024):** Surpassed 1,012 ExaFlops on the Top500 list, enabling foundation-model-scale AI for scientific discovery and simulation in the USA.
- **Alps Supercomputer (CSCS, 2024):** A pre-exascale system delivering approximately 270 PetaFlops, supporting climate modeling, physics simulations, and AI-enhanced numerical methods in Switzerland.
- **Isambard-AI (UK, 2024):** An AI-optimized leadership system targeting approximately 21 ExaFlops (FP8 AI performance), designed for LLM training, healthcare, and engineering applications in the United Kingdom.
- **El Capitan Super Computer (LLNL, 2024):** Deployed at Lawrence Livermore National Laboratory in USA, achieving approximately 1,742 exaflops and becoming the fastest system globally, with applications in national security, high-fidelity simulation, and classified AI workloads.
- **Jupiter Supercomputer (EuroHPC, 2025b):** Europe's first verified exascale system, supporting sovereign AI research and large-scale scientific workloads in Germany.
- **LUMI (EuroHPC, 2025b):** A top-tier European HPC system delivering approximately 0.375 exaflops for climate science and AI research in Finland.
- **Sunway OceanLight (NRCPC, 2025):** Ranked among the global top five systems, emphasizing indigenous processor development, aerospace research, and national technological independence in China.
- **ABCI 3.0 (AIST, 2025):** A next-generation AI infrastructure delivering multi-exaflop performance for large-scale AI, robotics, and trustworthy AI research in Japan.

- **Huawei Atlas 950 SuperCluster (Huawei, 2025):** Announced as one of the largest AI superclusters in China, targeting extreme-scale AI training and inference for industrial and societal applications.

Collectively, these developments reflect national strategies aimed at achieving leadership in AI, securing economic advantages, and enabling AI-driven transformation across defense, healthcare, education, energy, media, and environmental modeling (Top500, 2023). Most critically, such infrastructures and centers enable the development and secure deployment of national LLMs and LMMs, reducing dependency risks while ensuring data sovereignty, robustness, and resilience against emerging vulnerabilities.

4. Energy Consumption of Large AI Models

Recent advances in processor and accelerator technologies have been driven predominantly by the rapid expansion of large AI models, particularly since the widespread adoption of GenAI technologies after November 2022 (Vaswani et al., 2017). Semiconductor innovations have advanced chip manufacturers to the forefront of global technology markets. However, the large-scale deployment and training of LLMs, LMMs, VLM, etc, increasingly require massive GPU resources and dedicated AI data centers. As a consequence, the energy consumption associated with these systems has emerged as one of the most critical challenges in the sustainable development of Large AI models and beyond.

The rapid expansion of AI has led to growing concerns regarding its energy consumption and environmental impact. Prominent industry leaders have emphasized that the widespread deployment of large-scale models may significantly increase global energy demand, a view that is increasingly supported by empirical studies. Beyond direct electricity consumption during model training and inference, AI systems also impose substantial indirect environmental costs. These include carbon emissions associated with power generation, the embodied energy required for semiconductor manufacturing, and the infrastructure demands of large-scale data centers. Additionally, while efforts to develop smaller and more efficient models (i.e., compact, mini, and nano-scale architectures) aim to mitigate resource usage, the overall growth in AI adoption continues to amplify sustainability challenges across the entire lifecycle of AI systems.

Empirical findings from recent literature illustrate the magnitude of this issue. According to *The New Yorker* (Kolbert, 2024), ChatGPT is estimated to consume approximately 0.5 megawatt-hours (MWh) of electricity per day to handle around 200 million user requests, representing an energy demand several orders of magnitude higher than that of an average household. Vries (2023) also reported that a single ChatGPT query in 2023 consumes approximately 2.9 watt-hours (Wh) of

energy, compared to approximately 0.3 Wh for a traditional Google search, indicating nearly a tenfold increase in per-query energy consumption. Goldman Sachs Research (2024) also reports that "a ChatGPT query needs nearly 10 times as much electricity to process as a Google search" and also estimates that data center power demand will grow 160% by 2030. At the infrastructure level, multiple analyses also support for this forecast a dramatic rise in AI-driven data center electricity demand. By 2030, global data center electricity usage is estimated to exceed 550TWh per year, while NVIDIA projects that AI-specific workloads alone may consume between 85 and 134TWh annually by 2027 (NVIDIA, 2024a).

Similarly, the BLOOM LLM required approximately 433 MWh of electricity during training on 1.6 TB of data. A 2022 lifecycle assessment estimated that BLOOM's training resulted in 24.7 tons of CO_2 emissions, with an additional 50.5 tons attributable to hardware manufacturing and operational overhead (Luccioni et al., 2024). Microsoft has further reported that the energy consumption associated with initial model training far exceeds that of deployment and incremental fine-tuning, underscoring training as the dominant contributor to AI-related energy demand (Microsoft, 2023).

Large-scale industry examples further illustrate this trend, with a number of examples such as MetaAI reduces the parameters of LLaMa less than 10 times compared to GPT-3, training its initial LLaMA model over 21 days. MetaAI reported to use 2,048 NVIDIA A100 (NVIDIA, 2020) GPUs, processing 1.4 trillion tokens and consuming over one million GPU-hours. Training LLaMA-3 required more than 16,000 NVIDIA H100 GPUs (NVIDIA, 2022), processing approximately 15 trillion tokens and consuming nearly 39.3 million GPU-hours. Comparable large-scale training efforts reported by Grok to train their models utilized on the order of 100,000 processors during training.

As stated in Dauner & Socher (2025), the accuracy-emission trade-off has shown that there is a direct correlation between model scale and performance. Larger models generally achieve higher accuracy but at the cost of significantly higher emissions. For example: Qwen 7B is highly efficient (27.7g CO_2 but has low accuracy (32.9%); Deepseek-R1 70B reaches 78.9% accuracy but emits a massive 2,042.4g CO_2; NVIDIA trains a 1.7 trillion-parameter model with 8,000 Hopper-based GPUs required approximately 15 MW of power, whereas the same workload executed on 2,000 Blackwell B200 GPUs could reduce consumption to approximately 4 MW, demonstrating the critical role of next-generation accelerator efficiency (NVIDIA, 2024a; 2024b).

As reported in (Li et al., 2025) AI, expanding workloads in data centers not only will increase demand in the electricity consumption of AI servers surpassing TWh but also water consumption reaching billion cubic meters in 2027. Furthermore, increasing AI workloads have indirect environmental consequences, including water consumption for data center cooling. The cumulative effect of continuous model training, inference,

and infrastructure scaling contributes significantly to the carbon footprint of modern data centers, reinforcing the need for energy-aware AI system design (Ren et al., 2024; Li et al., 2025).

5. Increasing Hardware Requirements and Mitigation Strategies

The analysis above demonstrates that training and deploying large-scale generative AI models is both computationally and economically expensive. Consequently, not all organizations can or should undertake full-scale large model training independently. While no universal solution has yet emerged, collective and collaborative approaches are increasingly recognized as viable mitigation strategies. Likely, several ecosystem-level initiatives aim to provide access to computational resources. For example, Andreessen Horowitz has established shared GPU infrastructures comprising over 20,000 NVIDIA H100 GPUs to support AI startups (NVIDIA, 2022). Similarly, the Andromeda Cluster provides access to more than 4,000 GPUs at reduced cost to companies within its portfolio. Major cloud providers have also introduced targeted support programs and offered GPU credits to companies of venture funds such as Index Ventures, while Microsoft Azure, Google, Meta, Amazon Web Services (AWS) provide subsidized or free GPU access to selected startups. Beyond raw compute access, platform-level solutions are also provided to improve efficiency and usability for simplifying the deployment and management of big data and AI clusters, enabling organizations to optimize resource utilization, delivering comprehensive AI ecosystems encompassing infrastructure, development tools, and security services, thereby lowering entry barriers for researchers, enterprises, and educational institutions. As far as I know from our students in the department at Gazi University, these companies provide GPU credits to a number of entrepreneurial students.

Collectively, these approaches highlight a shift toward shared infrastructure, cloud-based optimization, and hardware–software co-design as essential strategies for addressing the escalating energy and hardware demands of AI systems.

6. Summary and Conclusion

This chapter highlights that the transformation of AI is fundamentally driven by the convergence of big data availability, advanced processor development, and the emergence of large-scale foundation models such as LLMs and LMMs**.** These three pillars collectively define the current trajectory of the AI ecosystem and represent one of the most significant technological shifts of the 21st century.

From a big data perspective, the exponential growth in the volume, velocity, and diversity of data has enabled the training of increasingly complex models. The availability of large-scale, multimodal datasets has been a critical enabler for foundation models, allowing them to generalize across tasks and domains. However,

this data-centric paradigm also introduces challenges related to data quality, labeling, privacy, and governance, which must be addressed to ensure reliable and responsible AI systems.

In parallel, processor development has evolved rapidly to meet the computational demands of modern AI workloads. The transition from general-purpose CPUs to specialized accelerators-such as GPUs, TPUs, NPUs, and domain-specific architectures-has significantly improved performance and efficiency. This evolution is further reinforced by advances in semiconductor technologies, including nanoscale fabrication, 3D packaging, and high-speed interconnects. At the system level, AI has driven the emergence of hyperscale data centers and exascale supercomputing infrastructures, creating a vertically integrated hardware–software ecosystem optimized for large-scale learning and inference.

Building on these foundations, the rise of LLMs and LMMs has redefined the capabilities of AI systems. These models leverage both massive datasets and advanced hardware to achieve unprecedented performance in language understanding, reasoning, and multimodal tasks. As a result, GenAI has become the dominant driver of innovation in both academia and industry, influencing semiconductor design, cloud architectures, and application ecosystems. At the same time, the widespread deployment of such models raises important concerns regarding energy consumption, environmental impact, security vulnerabilities, and equitable access to advanced AI technologies.

Looking forward, addressing these challenges requires a holistic approach centered on future solutions. Sustainable, responsible and trustable AI developments will depend on multiple complementary strategies, including:

- improving algorithmic efficiency and reducing model complexity,
- advancing hardware–software co-design for optimized performance per watt,
- integrating renewable energy sources and energy-aware computing paradigms,
- exploring emerging technologies such as neuromorphic, photonic, and analog computing,
- and establishing robust policy, governance, and security frameworks.

Furthermore, ensuring global inclusivity in AI development necessitates investments in national infrastructure, access to advanced processors, and the development of localized LLM/LMM ecosystems tailored to regional needs. Modern AI systems also face a range of security and robustness challenges, including model extraction, adversarial attacks, data poisoning, privacy leakage, prompt injection, and other vulnerabilities. When deployed in critical applications, these systems require formal

verification, rigorous safety guarantees, and comprehensive fault-tolerant architectures to mitigate risks from hardware failures, software bugs, and model errors.

Türkiye's growing engagement in semiconductor research, microfabrication facilities, and defense-oriented chip initiatives, implementing a national AI strategy and action plan, building up its own local AI models with local data and GPU resources reflects the national capacity-building interest for AI-enabling technologies. These dynamics underscore several strategic priorities such as focusing on establishing robust data center infrastructure, securing sufficient access to advanced GPUs and other accelerators, and developing a national LLM/LMM for general-purpose and sector-specific applications to foster technological sovereignty.

In conclusion, the future of AI will be shaped not by a single dimension, but by the synergistic interaction between big data, next-generation processors, and scalable foundation models, supported by sustainable and secure technological frameworks. Successfully balancing performance, efficiency, security, transparency, responsibility, accessibility, and sustainability will be critical in defining the next era of AI.

References

AMD. (2024a). AMD Ryzen AI 300 series processors. https://www.amd.com/en/products/processors/consumer/ryzen-ai.html

AMD. (2024b). AMD Ryzen 9000 series desktop processors. https://www.amd.com/en/products/processors/desktops/ryzen.html

AMD. (2024c). AMD XDNA architecture. https://www.amd.com/en/technologies/xdna.html

Argonne Leadership Computing Facility. (2024). Aurora exascale supercomputer. Argonne National Laboratory. https://www.alcf.anl.gov/aurora

Artificial Intelligence Bridging Cloud Infrastructure. (2025). ABCI 3.0: AI Bridging Cloud Infrastructure. https://abci.ai/

BrainChip. (2025). AKD1000 Akida system-on-chip product brief (Version 2.3). https://brainchip.com/

Chen, Z., et al. (2024). Evolution and prospects of foundation models: From large language models to large multimodal models. *Computers, Materials & Continua, 80*(2), 1753–1808. https://doi.org/10.32604/cmc.2024.052618

CSCS. (2024). Alps supercomputer system. https://www.cscs.ch/computers/alps/

Dauner, M., & Socher, G. (2025). Energy costs of communicating with AI. *Frontiers in Communication, 10*. https://doi.org/10.3389/fcomm.2025.1572947

DEEPX. (2024). On-device AI accelerator solutions. https://www.deepx.ai/

U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research. (2024). Discovery: Next-generation AI-native supercomputer program. https://www.energy.gov/science/ascr

EuroHPC Joint Undertaking. (2025a). JUPITER exascale supercomputer. https://eurohpc-ju.europa.eu/supercomputers/jupiter_en

EuroHPC Joint Undertaking. (2025b). LUMI supercomputer. https://www.lumisupercomputer.eu/

Genkina, D. (2024, May 8). Brain-inspired computer approaches brain-like size: Gestalt design yields flexibility and neuromorphic computational scale. *IEEE Spectrum.* https://spectrum.ieee.org/neuromorphic-computing-spinnaker2

Global Growth Insights. (2025). Neuromorphic chip market size – Forecast to 2035. https://www.globalgrowthinsights.com/market-reports/neuromorphic-chip-market-100008

Goldman Sachs Research. (2024). AI is poised to drive 160% increase in data center power demand. https://www.goldmansachs.com/insights/articles/ai-poised-to-drive-160-increase-in-power-demand

Google. (2017, May 12). An in-depth look at Google's first Tensor Processing Unit (TPU). Google Cloud Blog. https://cloud.google.com/blog/products/ai-machine-learning/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

Google Cloud. (2024). Cloud TPU. https://cloud.google.com/tpu

Stanford Institute for Human-Centered Artificial Intelligence. (2025). AI Index report 2025. Stanford University. https://hai.stanford.edu/ai-index/2025-ai-index-report

Huawei. (2025). Atlas 950 AI SuperCluster architecture white paper. *Huawei Cloud.* https://e.huawei.com/

Intel. (2021). Taking neuromorphic computing with Loihi 2 to the next level (Technology Brief).

https://download.intel.com/newsroom/2021/newtechnologies/neuromorphic-computing-loihi-2-brief.pdf

Kolbert, E. (2024, March 9). The obscene energy demands of A.I. The New Yorker. https://www.newyorker.com/news/daily-comment/the-obscene-energy-demands-of-ai

Li, P. F., Yang, J. Y., Islam, M. A., & Ren, S. (2025). Making AI less "thirsty": Uncovering and addressing the secret water footprint of AI models. *Communications of the ACM, 68*(7), 54–61. https://doi.org/10.1145/3724499

Lawrence Livermore National Laboratory. (2024). El Capitan supercomputer. https://computing.llnl.gov/el-capitan

Luccioni, A. S., Viguier, S., & Ligozat, A. L. (2024). Estimating the carbon footprint of BLOOM, a 176B parameter language model. *Journal of Machine Learning Research, 24*(1). https://jmlr.org/papers/v24/23-0069.html

Market Research Intellect. (2025). Photonic AI chip market size, share & trends analysis report 2025–2033. https://www.marketresearchintellect.com/

Middle East Technical University Micro-Electro-Mechanical Systems Research and Application Center. (2026). METU MEMS. https://mems.metu.edu.tr/en

Microsoft. (2023a). Microsoft Azure Maia. Microsoft Azure Blog. https://azure.microsoft.com/en-us/blog/

Microsoft. (2023b). Environmental sustainability. Microsoft. https://www.microsoft.com/en-us/sustainability

Miscuglio, M., & Sorger, V. J. (2020). Photonic tensor cores for machine learning. *Applied Physics Reviews, 7*(3), 031404. https://doi.org/10.1063/5.0001942

National Research Center of Parallel Computer Engineering and Technology. (2025). Sunway OceanLight supercomputer. TOP500. https://www.top500.org/system/179973/

NVIDIA. (2017). NVIDIA Tesla V100 GPU architecture (White Paper WP-08608-001 v1.1). https://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf

NVIDIA. (2020). NVIDIA A100 Tensor Core GPU architecture (White Paper v1.0). https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf

NVIDIA. (2022). NVIDIA H100 Tensor Core GPU architecture (White Paper v1.0). https://resources.nvidia.com/en-us-tensor-core

NVIDIA. (2023). NVIDIA H200 Tensor Core GPU (Product Brief). https://www.nvidia.com/en-us/data-center/h200/

NVIDIA. (2024a). NVIDIA Blackwell platform. https://www.nvidia.com/en-us/data-center/technologies/blackwell-architecture/

NVIDIA. (2024b). NVIDIA GB200 Grace Blackwell Superchip. https://www.nvidia.com/en-us/data-center/grace-blackwell-superchip/

NVIDIA. (2024c). Semiconductor design with large language models. *NVIDIA Technical Blog*. https://developer.nvidia.com/blog/

NVIDIA. (2024d). Performance and energy efficiency comparison: Hopper vs. Blackwell architecture. *NVIDIA.* https://www.nvidia.com/en-us/data-center/

Open Neuromorphic. (2025). Neuromorphic hardware guide. https://openneuromorphic.org/neuromorphic-computing/hardware/

Oak Ridge National Laboratory. (2023). Frontier: The world's first exascale supercomputer. https://www.olcf.ornl.gov/frontier/

Ren, H., et al. (2024). ChipNeMo: Domain-adapted LLMs for chip design (arXiv:2311.00176v5). arXiv. https://arxiv.org/abs/2311.00176

Ren, S., Li, P., & Zhang, M. (2023). The carbon footprint of data centers and AI: Challenges and solutions. *IEEE Access, 11*, 45632–45648. https://doi.org/10.1109/ACCESS.2023.3274521

Rebellions. (2023). ATOM: AI accelerator for data centers. https://www.rebellions.ai/

Rebellions. (2024). ION SSD: AI inference accelerator. https://www.rebellions.ai/products

Shen, Y., Harris, N. C., Skirlo, S., et al. (2017). Deep learning with coherent nanophotonic circuits. Nature Photonics, 11(7), 441–446. https://doi.org/10.1038/nphoton.2017.93

Sparkes, M. (2025, October 29). Analogue computers could train AI 1000 times faster and cut energy use. New Scientist. https://www.newscientist.com/article/ 2500442-analogue-computers-could-train-ai-1000-times-faster-and-cut-energy-use/

Stackpole, B. (2025, August 19). New MIT report captures state of quantum computing. MIT Sloan Management Review. https://mitsloan.mit.edu/ideas-made-to-matter/new-mit-report-captures-state-quantum-computing

TOP500. (2025). TOP500 supercomputer list. https://www.top500.org/

TSMC. (2024a). Technology leadership. https://www.tsmc.com/english/dedicatedFoundry/technology

TSMC. (2024b). 2 nm process technology. https://www.tsmc.com/english/dedicatedFoundry/technology/logic/l_2nm

TÜBİTAK BİLGEM. (2026). Yarı iletken teknolojileri. https://bilgem.tubitak.gov.tr/yari-iletken-teknolojileri/

University of Bristol, & UK Research and Innovation. (2024). *Isambard-AI: UK national AI supercomputer.* https://www.bristol.ac.uk/acrc/isambard/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008. https://proceedings.neurips.cc/paper/ 2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

de Vries, A. (2023). The growing energy footprint of artificial intelligence. *Joule, 7*(10), 2191–2194. https://doi.org/10.1016/j.joule.2023.09.004

Xue, Y., Liu, Y., Nai, L., & Huang, J. (2024). Hardware-assisted virtualization of neural processing units for cloud platforms. In *Proceedings of the 57th IEEE/ACM International Symposium on Microarchitecture* (MICRO) (pp. 1–16). IEEE. https://doi.org/10.1109/MICRO61859.2024.00011

Yao, M., Richter, O., Zhao, G., et al. (2024). Spike-based dynamic computing with asynchronous sensing-computing neuromorphic chip. *Nature Communications, 15*, 4464. https://doi.org/10.1038/s41467-024-47811-6

Zuo, P., & Sun, Z. (2025). RRAM-based analog matrix computing for massive MIMO signal processing: A review. arXiv. https://doi.org/10.48550/arXiv.2512.04365

About the Author

Prof. Dr. Şeref SAĞIROĞLU
President of AI Ecosystem Association | ss[at]gazi.edu.tr
ORCID: 0000-0003-0805-5818

Prof. Dr. Şeref Sağıroğlu is President of the AI Ecosystem Association of Türkiye since 2025. He continues his academic career as a professor in Software Engineering at Gazi University Computer Engineering Department. Prof. Sağıroğlu has an outstanding academic having more than 150 SCI/SSCI indexed articles, 150 national and International Indexed Articles, more than 250 National and International Conference and Symposium Articles, author and/or Editor of More than 20 Books, editors of Open Source Books (Cyber Security and Defense Book Series; Digital Literature; Artificial Intelligence and Big Data Book Series; Open and Big Data Series), 15 Patents (7 pending), National and International Projects. Sağıroğlu has organized more than 50 national and international events as chairman, co-chairman. He has also been founding members of Information Security Association, Turkish Science Research Foundation, The Foundation of the People Caring for the Future, Information Security Engineering Graduate and Post Graduate Programs of Gazi University, Big Data Analytics, Security and Privacy Program of Gazi University, Software Engineering Undergraduate Program. Sağıroğlu has/had such duties as Dean of Engineering Faculty of Gazi University, Member of ETSI as an Observer, President and Executive Committee Member of Information Security Association, President and Member of Turkish Science and Research Foundation, Dean of Graduation School of Science and Technology at Gazi University, Founding Director of Gazi AI Center (2021-2024), Head of Computer Engineering Department, Gazi University, Erciyes University. Member of IEEE Biometric Task Force, Senior Member of IEEE, President of IEEE Blockchain Türkiye Working Group, President of IPv6 Forum Türkiye, International Journal of Information Security Science, International Journal of Information Security Engineering, and CyberMag, General Director of FutureTech, Member of Cyber Security Group of Higher Education Council of Türkiye, Head or established new Departments of Software Engineering program, Information Security Engineering Program, Smart Grid Program, Big Data Analytics, Security and Privacy. Sağıroğlu also has voluntarily been involved in many social projects and carried out projects in Scientific Research Projects such as TUBITAK, T.C. Digital Transformation Office, TUSEB, European Union, Sectors, University BAP Unit. Prof. Sağıroğlu has delivered as invited or keynote speakers more than 500 seminars, talks, conferences at universities, schools, sectors, TV and Radio Programs, institutions and organizations in the topics of Information Security; Big and Open Data; Cyber Security and Defense; Artificial Intelligence; Artificial Neural Networks and Its Applications; Generative AI, Software Engineering; Privacy; Biometrics; Innovation Culture Creation; IPv6, Big Data Analytics; Robotics; 5G; Industry 4.0; etc.