Advanced Gen AI & LLM Foundations and Applications

– Paving the way to a more powerful and diverse ML –*

Prof. Dr. V. David Sánchez A., Ph.D.
Brilliant Brains, Palo Alto, California
December 2023

Abstract

The advances in the field of Artificial Intelligence (AI) in less than 70 yrs after the Darmouth first U.S. ”workshop” on Artificial Intelligence [2, 3] in 1956 organized by J. McCarthy have been substantial. To that point, A.M. Turing had already defined his ”imitation game” [4], now known as the Turing test, initially considering the question: ”Can machines think?”. If a machine, e.g., a computer can win that game, it can be called intelligent, so his key idea. It replaced the original question by the question of whether a machine can act indistinguishably from the way a human acts. In other words, whether a human and an intelligent machine act in the same way while playing that game. A type of reverse test is the popular, current captcha test to determine whether an online user is really a human or a machine (bot).

For all understanding of the early beginning's efforts, those formal definitions of intelligence were/are fully insufficient even compared to any layman's understanding of the word intelligence. Basic understanding could even lead immediately to the clear answer to long-standing questions whether at least some animals are intelligent since computers cannot see ”so well” (even in the meantime) while those animals can, etc. I was delighted to conceive, design and deliver the first operational, mission-critical, real-time, parallel distributed supercomputer capable of performing robotic vision and control in history as an integral system operated even fully robotic-autonomously (100%) in a flown NASA/ESA/DLR Spaceshuttle/Spacelab mission [5], see Figure 1. According to J.W. Freeman's life-time modeling research of the biological cortex: ”The (David's) associated areas of scientific research have formed the baseline within brain research towards a more complete understanding of the operation of biological brains including the human one” [6].

On the other hand, multiple ambitious predictions about AI performance could not fulfill expectations, so, e.g., M. Minsky's in November 1970 according to Life magazine, not a scientific publication, but still: ”In from three to eight years we will have a machine with the general intelligence of an average human being. I mean a machine that will be able to read Shakespeare, grease a car, play office politics, tell a joke, have a fight. At that point the machine will begin to educate itself with fantastic speed. In a few months it will be at genius level and a few months after that its powers will be incalculable.” [7]. M. Minsky and S. Papert had previously published ”The Perceptrons: An Introduction to Computational Geometry” with the main subject being the perceptron, a type of basic artificial neural network, acknowledging its strengths, but also its limitations, which caused at least partly a redirection of research. Today, more than a half a century after those predictions, nothing

*This abstract has been granted permission for public release. The author has among many others a 15-year tenure as Chief Scientist and Editor-in-Chief of an Elsevier Science's artificial intelligence, machine learning journal [1] and is the youngest ”Nobel”-Prize in Engineering (IEEE Fellow) in history worldwide with the following mention ”for leadership in neural and parallel computation, and pioneering contributions to autonomous space robots”. The author introduced advanced self-building AI/ML architectures and robust learning methods which could also process sequences back in the 1980's before transformers were introduced in 2017 to ”replace” RNNs and CNNs for that purpose and led to the development of the family of Generative Pre-trained Transformers (GPTs) and Large Language Models (LLMs) in Gen AI.

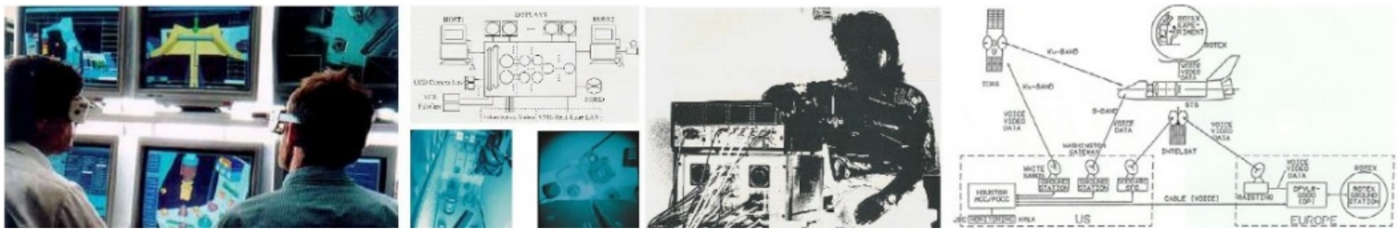


Figure 1: NASA STS-55 Spacelab Mission – First space operational computer, robot vision system, teleoperated, shared-controlled, and fully autonomous. [13]

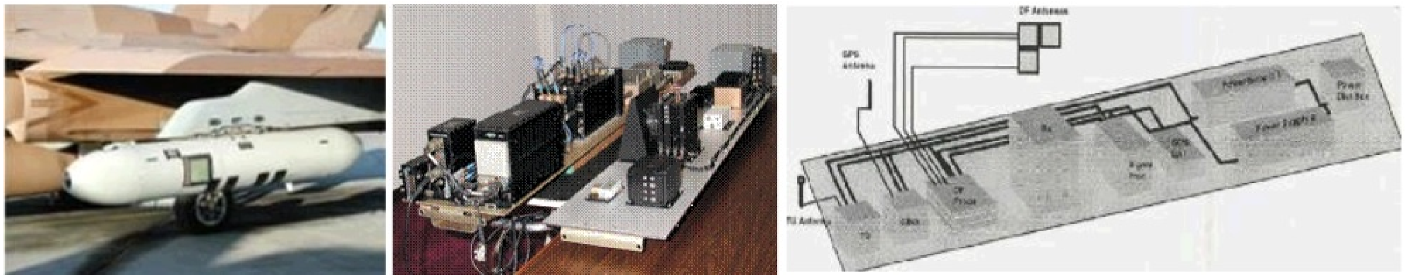


Figure 2: DoD advanced avionics with intelligent built-in sensor system – Mission-critical use of neurochips integrated with CPUs, DSPs, GPS/INS among others.

the like has been achieved. It is certain that the complexity of the essence and details of intelligent computation has been by far continuously underestimated.

After being heavily involved in ”AI for VLSI”, i.e., I designed, coded, and deployed an AI-tool at the Karlsruhe Institute of Technology (KIT) to design ASICs which we then used at Siemens AG to build custom processors for automation, I founded with around 15 experts a committee on ”VLSI for AI” in 1985 in Berlin, Germany. I have been committed to advancing AI and ML since the early 1980’s including pioneering breakthrough applications, industrial and in research [9], advanced learning / training methods [10] as well as the early use of neurochips, cf. Figure 2, and advanced development environments [11]. Recent advances in machine learning have been reported in [12], with topic intersection in research and applications using big data DevSecOps in [13], in computational data science [14], and key government regulatory issues of AI technology in [15] providing some insight into AI supercomputers for the Gen AI era. The provided citations exemplify relevant developments and are by no means meant to be exhaustive.

Figure 3 relates in (a) underlying Gen AI disciplines to one another including Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), Generative AI (Gen AI), and Large Language Models (LLMs), shows in (b) deep learning model types: discriminative and generative, and in (c) that Generative AI and Large Language Models are forms of deep learning. Discriminative DL models are used to classify or predict and are typically trained on datasets of labeled data (supervised learning). Generative DL models generate new data similar to the data they were trained on. From a machine learning (ML) perspective called deep learning (DL), knowledge is gathered from experience (data) without human intervention and represented by a hierarchy of concepts, which internally, i.e., in the machine, is structured via a hierarchy graph composed of a number of layers deep. For a broad range of basic deep learning topics consult, e.g., [16]. The deep learning technological and scientific impact has been addressed, e.g., in [17]. Some key mathematical foundations of machine learning and central machine learning problems can be found, e.g., in [18]. Some basic applications and associated statistical learning techniques including classification, regression, resampling, model selection, tree-based methods, survival analysis, and multiple testing can be found, e.g., in [19]. Introductory topics of GenAI and LLMs have been or are being prepared, e.g., in [20] and in [21] from a business perspectve, just to mention a few of a vast list of publications about this relatively new subject. LLM topics like an introduction to Large Language Models (LLMs) including

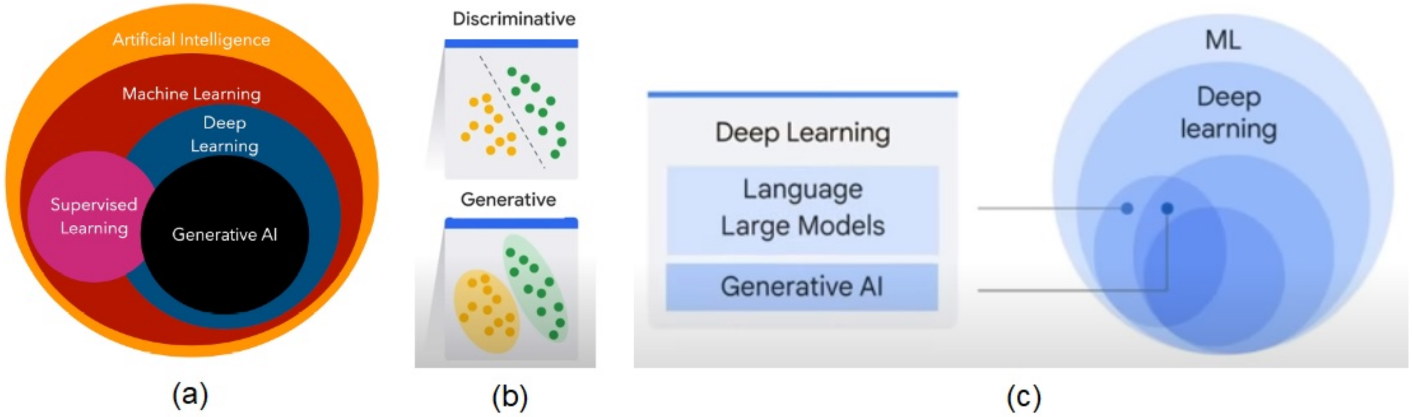


Figure 3: Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), Generative AI (Gen AI), Large Language Model (LLM): (a) Gen AI related disciplines, (b) DL model types, (c) Gen AI and LLMs are forms of DL [CMU, Google]

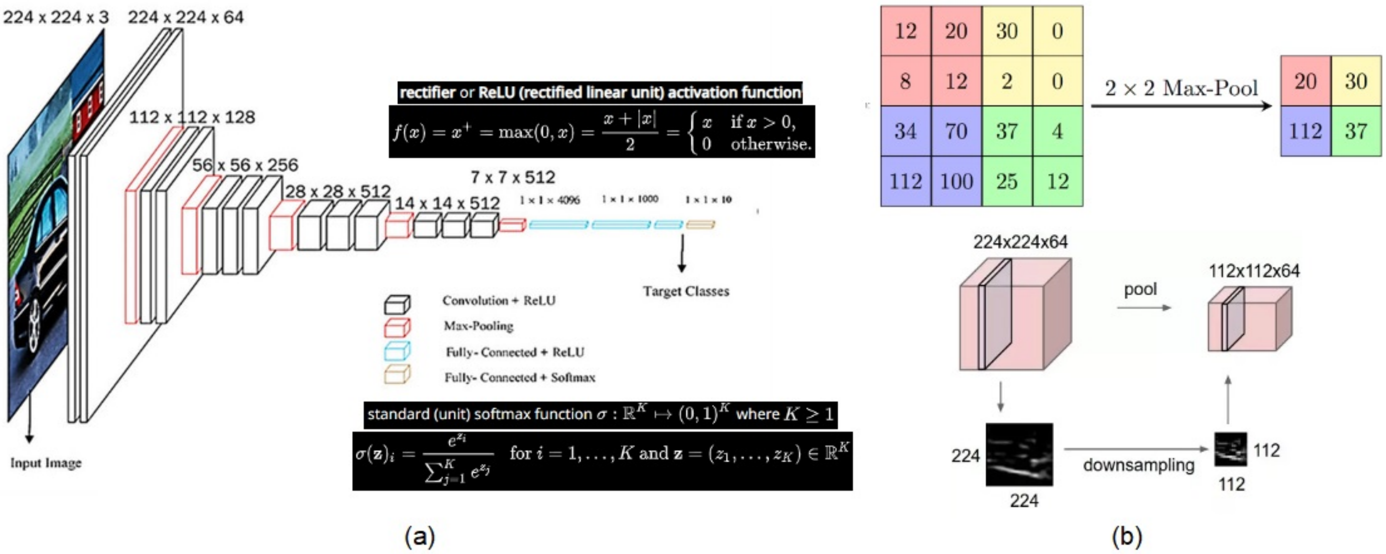

224 x 224 x 3
224 x 224 x 64
112 x 112 x 128
56 x 56 x 256
28 x 28 x 512
14 x 14 x 512
7 x 7 x 512
1 × 1 × 4096
1 × 1 × 1000
1 × 1 × 10
Target Classes
Input Image
Convolution + ReLU
Max-Pooling
Fully-Connected + ReLU
Fully-Connected + Softmax
rectifier or ReLU (rectified linear unit) activation function
$f(x) = x^+ = \max(0, x) = \frac{x + |x|}{2} = \begin{cases} x & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases}$
standard (unit) softmax function $\sigma : \mathbb{R}^K \mapsto (0,1)^K$ where $K \geq 1$
$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}}$ for $i = 1, \ldots, K$ and $\mathbf{z} = (z_1, \ldots, z_K) \in \mathbb{R}^K$
(a)
12 20 30 0
8 12 2 0
34 70 37 4
112 100 25 12
2 × 2 Max-Pool
20 30
112 37
224x224x64
pool
112x112x64
224
224
downsampling
112
112
(b)

Figure 4: Convolutional Neural Network (CNN) example [24]: (a) VGG-16 architecture with 16 layers, ReLU, Softmax (b) Max-Pooling

an overview of LLMs, semantic search with LLMs, prompt engineering first steps; getting the most out of LLMs including optimizing LLMs with customized fine-tuning, advanced prompt engineering, customizing embeddings and model architectures; and advanced LLM usage: beyond foundation models, open-source LLM fine-tuning, moving LLMs into production are presented, e.g., in [22].

In computer vision, a frequent way to extract features from images is by using convolution operators. In machine learning, an appealing way to perform image classification is by using Convolutional Neural Networks (CNNs) [23], typically composed of an upstream feature extractor followed by a downstream classifier. As an example, in [24] simple Convolutional Neural Network models were presented with 16 and 19 layers depth by the Oxford's Visual Geometry Group (VGG), in particular with greater depth than with AlexNet [25]. Other deep CNN architectures can be found, e.g., in [12]. Figure 4 shows the VGG-16 network with 16 layers including built-in functions and operations utilized in the architecture: ReLU, Max-Pooling, Softmax. The depth of Deep Neural Networks (DNNs), and in particular of Convolutional Neural Networks (CNNs) accounts for the higher accuracy of the results obtained and a hierarchical and modular representation within the architecture with subsequently higher levels of abstraction, in the case of image classification going from a layer of pixels to a layer of objects going through layers of edges among others.

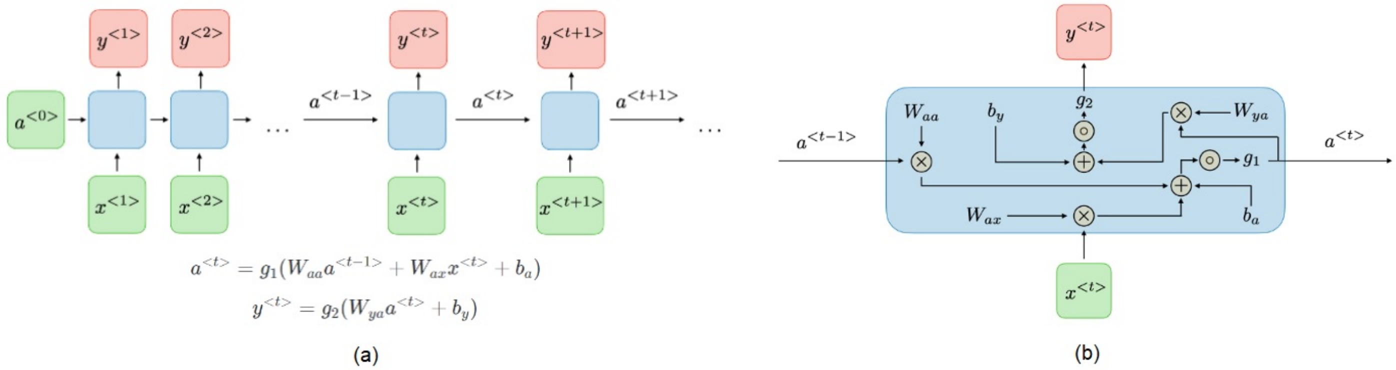


Figure 5: Recurrent Neural Network (RNN) to process sequential data {time series, natural language} [Stanford]

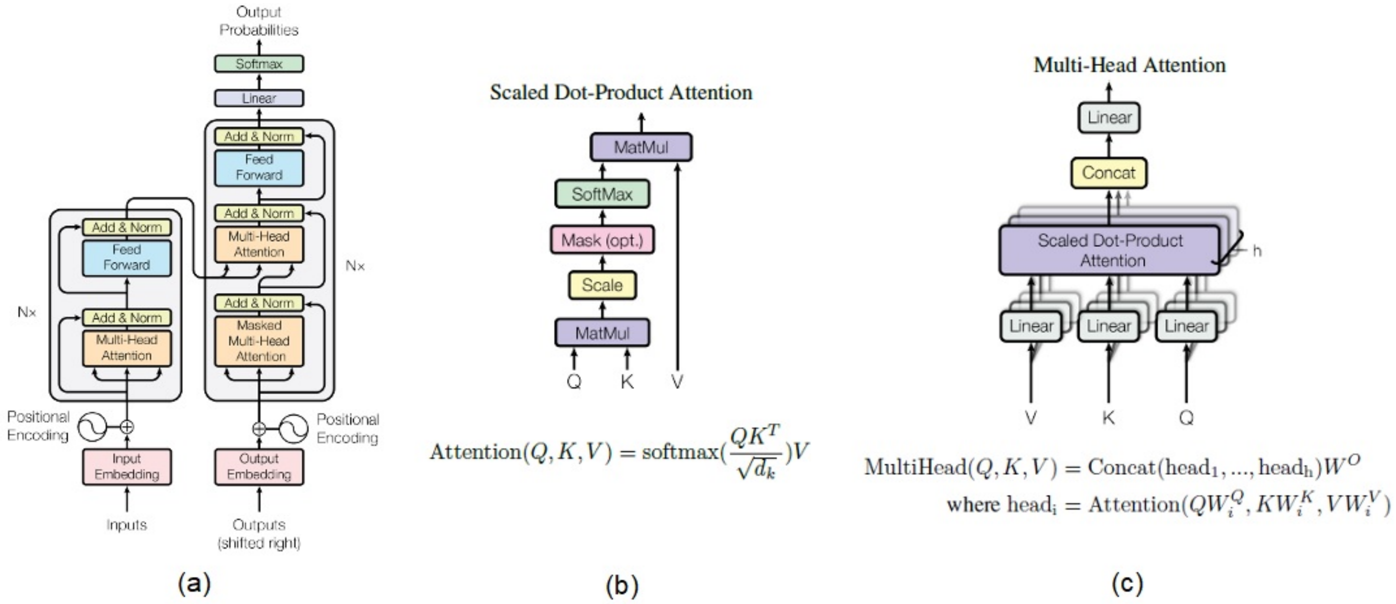
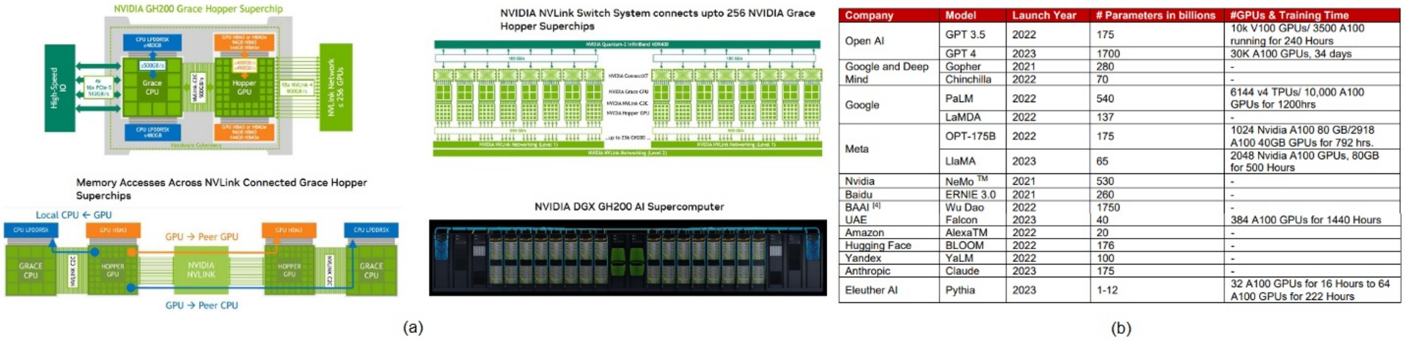


Figure 6: Transformer network architecture based solely on attention mechanisms [29]

Figure 5 shows in (a) the traditional recurrent neural network (RNN) architecture with hidden states, which allows past outputs to be used as inputs and the processing of sequential data like time series and natural language, its activation $a^{<t>}$ and output $y^{<t>}$ for t, the timestep, and input $x^{<t>}$; and in (b) the temporarily shared coefficients $W_{ax}, W_{aa}, W_{ya}, b_a, b_y$ and the activation functions g_1 and g_2. The gradient-based training technique for recurrent networks Backpropagation through time (BPTT) was introduced in [26]. Generalized BPTT training methods for recurrent neural networks are presented, e.g., in [27], whose local optima issues are more challenging than with feed-forward neural networks. Another sequence learning method that addresses the RNN's vanishing gradient issue can be found, e.g., in [28], called the Long Short-Term Memory (LSTM).

A self-attention-based sequence transduction model called Transformer was introduced in [29]. The underlying attention mechanism relates positions of a single sequence to compute the sequence representation, without using neither Recurrent Neural Networks (RNNs) nor Convolutional Neural Networks (CNNs) to connect the encoder and decoder parts, and allowing for a higher degree of paralellizable computation. An attention function is a mapping of a query and a set of key-value pairs to an output, where each of these terms are vector-valued. The weighted values sum is assigned to the output whereas a weight is determined using a compatibility function of the query with the corresponding key. Figure 6 shows in (a) the transformer model architecture and in (b) the scaled dot-product and in (c) the multi-head attention functions. With the multi-head attention function,


NVIDIA GH200 Grace Hopper Superchip
High-Speed IO
4x 16x PCIe-5 512GB/s
CPU LPDDR5X ≤480GB
≤500GB/s
Grace CPU
NVLink C2C 900GB/s
GPU HBM3 or HBM3e 96GB HBM3 144GB HBM3e
≤4000GB/s ≤4900GB/s
Hopper GPU
18x NVLink 4 900GB/s
NVLink Network ≤ 256 GPUs
Hardware Coherency
NVIDIA NVLink Switch System connects upto 256 NVIDIA Grace Hopper Superchips
NVIDIA Quantum-2 InfiniBand NDR400
NVIDIA ConnectX7
NVIDIA Grace CPU
NVIDIA NVLink C2C
NVIDIA Hopper GPU
...up to 256 GH200 ...
NVIDIA NVLink Networking (Level 1)
NVIDIA NVLink Networking (Level 2)
Memory Accesses Across NVLink Connected Grace Hopper Superchips
Local CPU ← GPU
CPU LPDDR5X
GPU HBM3
GPU → Peer GPU
GRACE CPU
NVLINK C2C
HOPPER GPU
NVIDIA NVLINK
GPU → Peer CPU
NVIDIA DGX GH200 AI Supercomputer
(a)
(b)

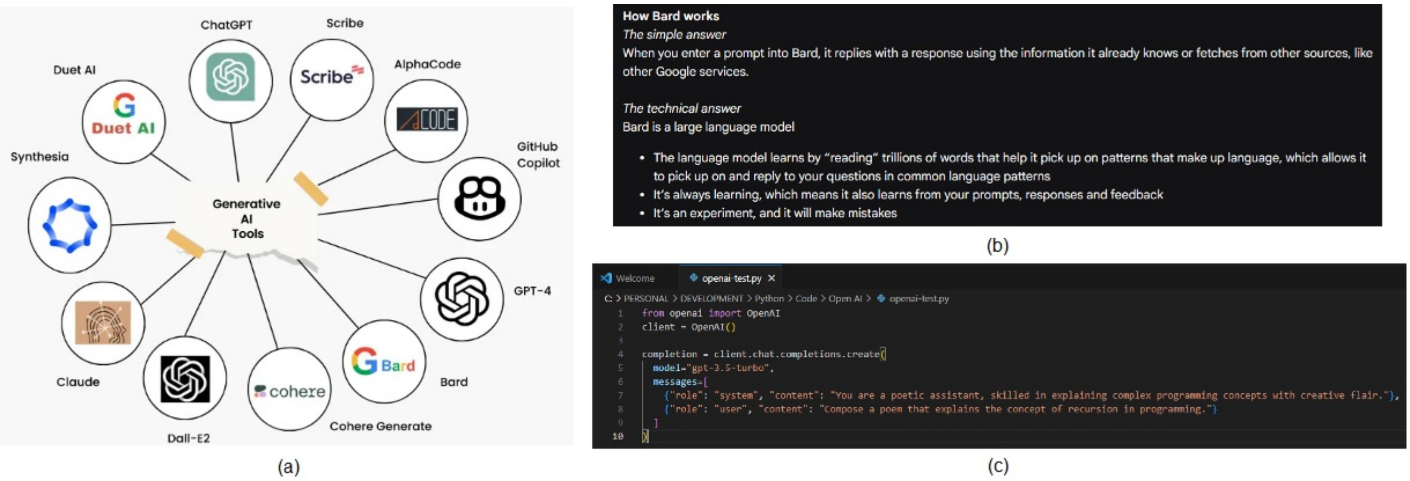
Company	Model	Launch Year	# Parameters in billions	#GPUs & Training Time
Open AI	GPT 3.5	2022	175	10k V100 GPUs/ 3500 A100 running for 240 Hours
	GPT 4	2023	1700	30K A100 GPUs, 34 days
Google and Deep Mind	Gopher	2021	280	-
	Chinchilla	2022	70	-
Google	PaLM	2022	540	6144 v4 TPUs/ 10,000 A100 GPUs for 1200hrs
	LaMDA	2022	137	-
Meta	OPT-175B	2022	175	1024 Nvidia A100 80 GB/2918 A100 40GB GPUs for 792 hrs.
	LlaMA	2023	65	2048 Nvidia A100 GPUs, 80GB for 500 Hours
Nvidia	NeMo TM	2021	530	-
Baidu	ERNIE 3.0	2021	260	-
BAAI [4]	Wu Dao	2022	1750	-
UAE	Falcon	2023	40	384 A100 GPUs for 1440 Hours
Amazon	AlexaTM	2022	20	-
Hugging Face	BLOOM	2022	176	-
Yandex	YaLM	2022	100	-
Anthropic	Claude	2023	175	-
Eleuther AI	Pythia	2023	1-12	32 A100 GPUs for 16 Hours to 64 A100 GPUs for 222 Hours

Figure 7: (a) Generative AI Supercomputer [Nvidia] [15] (b) Model Parameter Size [33]

the transformer model can jointly attend to information from different representation subspaces at different positions.

This basic architecture led to the development of complex LLMs all the way to OpenAI's fourth generation of GPT (Generative Pre-trained Transformer) foundation models. GPT-4, for example, is a multi-modal (text, image) LLM made publicly available via ChatGPT Plus and OpenAI's API [30]. The pre-trained model uses public and third-party providers' licensed data, which is then fine-tuned using reinforcement learning feedback from humans (RLFH) and AI for human alignment and policy compliance. Rising capabilities and implications of an early version of the GPT-4 model are discussed, e.g., in [31], providing a phenomenological study over a range of tasks and domains given a loose understanding of reasoning, planning, and learning from experience as well as highlighting the remarkable capabilities and challenges of GPT-4. Some of the current AI technology developments with major impact obviously include the deployment of next generation hardware too, e.g., the Nvidia's DGX GH200, an AI supercomputer for the Generative AI era [32, 15]. AI hardware has substantially accounted for the increase in the size of the Gen AI models, i.e., the number of model parameters, see, e.g., [33]. Figure 7 shows in (a) the DGX GH200 Gen AI supercomputer and in (b) the increase in foundation model parameter number up to 1.7 trillion for GPT-4.

Multiple areas of research and improvement include, e.g., the fine-tuning of large language models in practice by training only a small set of parameters. That small set could be either a subset of the current parameters or a set of new parameters not present inside the model yet. A parameter-efficient fine-tuning (PEFT) method taxonomy is presented, e.g., in [34]. Zero-shot, one-shot, and few-shot learning are used when feasible and the availability of training data is insufficient. Max-likelihod-trained LLMs appear to perform surprisingly well for diverse tasks in a zero-shot setting and without supervision when trained using a sufficiently varied text corpus [35]. An approach to multi-tasking model development without negative interference is making use of modular deep learning, see, e.g., [36]. Tools have evolved since the Eliza program developed by J. Weizenbaum made some basic natural language conversation between human and machine possible [37]. Figure 8 shows in (a) the landscape of some of the key Generative AI tools today including AlphaCode, Bard, ChatGPT, Claude, Cohere Generate, Dall-E2, Duet AI, GitHub Copilot, GPT-4, Scribe, and Synthesia, in (b) as an example of a Gen AI tool, how Google's Bard operates and in (c) as another example of a Gen AI tool, the way how to send an Open AI API request using the custom Python library provided after the API key has been set up. Open AI's GPT-4 can be seen in brief as a multimodal (text, image) Gen AI tool for content creation, answering relatively complex questions, and could be used in marketing applications among others; Google's Bard as a conversational Gen AI tool to brainstorm ideas, spark creativity, and accelerate productivity, which could be used for dialog-based applications and customer service. Figure 9 (a) shows the Bard's response to a mathematical task, i.e., when entering the prompt: "Suppose g(x) =f ^{-1}(x); g(0) = 5; g(4) = 7; g(3) = 2; g(7) = 9; g(9) = 6 what is f(f(f(6)))?". Bard solves this task correctly in three steps: (1) f(6) = 9, (2) f(f(6)) = f(9) = 7, and finally (3) f(f(f(6))) = f(f(9)) = f(7) = 4. Btw., the result is the same as with GPT-4 and the "reasoning" displayed pretty similar. Figure 9 (b) shows the Bard's response to a coding task, i.e., when entering



(a)

How Bard works

The simple answer

When you enter a prompt into Bard, it replies with a response using the information it already knows or fetches from other sources, like other Google services.

The technical answer

Bard is a large language model

- The language model learns by "reading" trillions of words that help it pick up on patterns that make up language, which allows it to pick up on and reply to your questions in common language patterns
- It's always learning, which means it also learns from your prompts, responses and feedback
- It's an experiment, and it will make mistakes

(b)

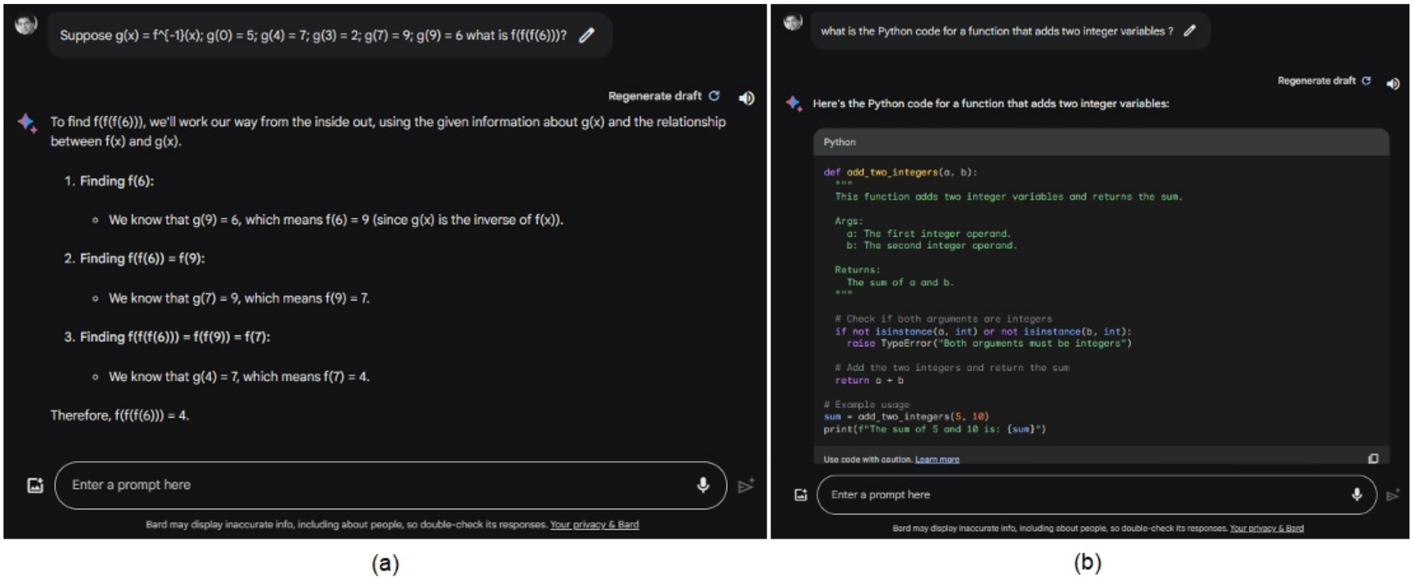
Welcome
openai-test.py
C: > PERSONAL > DEVELOPMENT > Python > Code > Open AI > openai-test.py

```
from openai import OpenAI
client = OpenAI()

completion = client.chat.completions.create(
  model="gpt-3.5-turbo",
  messages=[
    {"role": "system", "content": "You are a poetic assistant, skilled in explaining complex programming concepts with creative flair."},
    {"role": "user", "content": "Compose a poem that explains the concept of recursion in programming."}
  ]
)
```

(c)

Figure 8: (a) Gen AI Tools (b) Google's Bard (c) Open AI's GPT-4 – API request



Suppose g(x) = f^{-1}(x); g(0) = 5; g(4) = 7; g(3) = 2; g(7) = 9; g(9) = 6 what is f(f(f(6)))?

To find f(f(f(6))), we'll work our way from the inside out, using the given information about g(x) and the relationship between f(x) and g(x).

1. Finding f(6):
   - We know that g(9) = 6, which means f(6) = 9 (since g(x) is the inverse of f(x)).
2. Finding f(f(6)) = f(9):
   - We know that g(7) = 9, which means f(9) = 7.
3. Finding f(f(f(6))) = f(f(9)) = f(7):
   - We know that g(4) = 7, which means f(7) = 4.

Therefore, f(f(f(6))) = 4.

Bard may display inaccurate info, including about people, so double-check its responses. Your privacy & Bard

(a)

what is the Python code for a function that adds two integer variables ?

Here's the Python code for a function that adds two integer variables:

```
def add_two_integers(a, b):
  """
  This function adds two integer variables and returns the sum.

  Args:
    a: The first integer operand.
    b: The second integer operand.

  Returns:
    The sum of a and b.
  """

  # Check if both arguments are integers
  if not isinstance(a, int) or not isinstance(b, int):
    raise TypeError("Both arguments must be integers")

  # Add the two integers and return the sum
  return a + b

# Example usage
sum = add_two_integers(5, 10)
print(f"The sum of 5 and 10 is: {sum}")
```

Use code with caution. Learn more

Bard may display inaccurate info, including about people, so double-check its responses. Your privacy & Bard

(b)

Figure 9: Bard generated responses: (a) mathematical task, (b) coding task.

the prompt: ”What is the Python code for a function that adds two integer values ?”. The result provided by Bard includes a usage example of the Python code generated.

Figure 10 (a) shows the image produced by LaTeX compiling after GPT-4 processes the prompt: ”Draw a unicorn in TiKZ”. The TikZ package is a complex and powerful tool to create graphic elements in LaTeX, producing vector graphics, e.g., technical illustrations and drawings, from a geometric/algebraic description. Figure 10 (b) displays the three versions of the GPT-4 results showing an increased degree of the drawings' sophistication after being queried three(3) times subsequently within approximately one month while the system was being refined. Figure 10 (c) shows on the left the famous oil on canvas called ”Komposition 8” by Wassily Kandinsky, credited to be one of the pioneers of abstraction in western art and on the right the image produced by the code that was generated entering the prompt: ”Produce javascript code which generates random images in the style of the painter Kandinsky”.

The transformative effect of these pre-trained, self-supervised foundation models called Large Language Models (LLMs) in natural language processing (NLP) using current Generative Artificial Intelligence (Gen AI) tools has been remarkable, mainly due to its potential adaptability via fine-tuning to a broad range of tasks. These advances have led to a relatively vast and fast US government initial reaction as the following examples show. The U.S. Department of Defense (DoD) has released

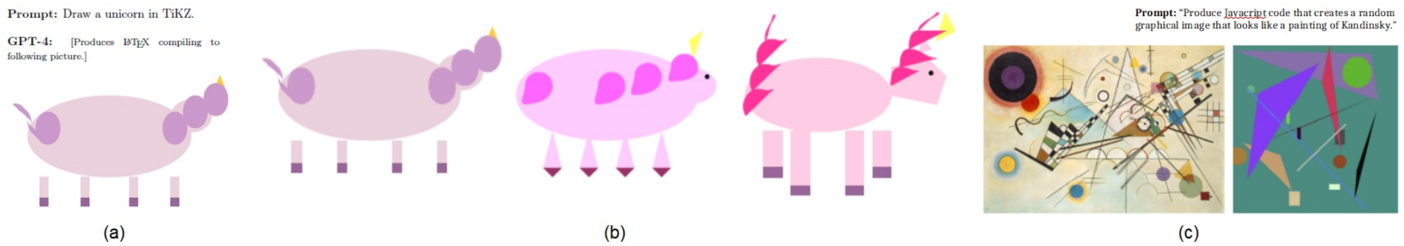

Prompt: Draw a unicorn in TiKZ.
GPT-4: [Produces LaTeX compiling to following picture.]
Prompt: "Produce Javacript code that creates a random graphical image that looks like a painting of Kandinsky."
(a)
(b)
(c)

Figure 10: GPT-4 generated images: (a) first image as GPT-4 query result to the prompt entered, (b) Three GPT-4 query results with increasing degree of image sophistication, (c) left: Wassily Kandinsky's Komposition 8, right: image produced by code generated by GPT-4 as response to the prompt shown [31]

a strategy that guides the strengthening of the organizational environment in which the DoD deploys data, analytics, and AI capabilities for enduring decision advantage [38] based on lessons learned while integrating analytics and AI applications and first-hand learning about their current benefits and limitations. The White House released at the end of October 2023 an executive order on the safe, secure, and trustworthy development and use of Artificial Intelligence [39, 40]. W.r.t. the federal use of AI, it acknowledges the ubiquity of Gen AI tools, and directs agencies to provide access with safeguards in place. Some of the key AI-related activities of the U.S. Congress (Senate & House) have been summarized in [15, 41]. To support AI and national security the Special Competitive Studies Project (SCSP) was formed in October 2021 composed of six panels with the following objectives: foreign policy, intelligence, defense, economy, society, and future tech platforms. In particular, with emphasis on Gen AI [42]. The entire area of generative Artificial Intelligence (Gen AI) and Large Language Models (LLMs) is reviewed and benchmarked in detail providing relevant topics of current research, advanced development tools, stacks and use cases as well as risk assessments while incorporating the associated technologies into vertical applications in multiple areas of science, industrial, and government application.

References

[1] V.D. Sánchez. *Neurocomputing 50th volume anniversary*. Neurocomputing, 50, ix, 2003.

[2] J. McCarthy et al. *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. August 31, 1955.

[3] J. Moor. *The Dartmouth College Artificial Intelligence Conference: The Next Fifty years*. AI Magazine, 27 [4], 87–89, 2006.

[4] A.M. Turing. *Computing Machinery and Intelligence*. Mind, 49, 433–460, 1950.

[5] NASA. *STS-55 Spacelab Mission*, April 26 to May 6, 1993.
https://www.nasa.gov/mission/sts-55/

[6] J.W. Freeman. *University of California, Berkeley*. Division of Neurobiology, April 2012.
https://profdrvdsaphd.lima-city.de/documents/Recommendation8.pdf

[7] B. Darrach *Meet Shaky, the first electronic person – The fascinating and fearsome reality of a machine with a mind of its own*. Life Magazine November 20, 1970, 58B–68.

[8] M. Minsky and S. Papert. *Perceptrons: An Introduction to Computational Geometry*. The MIT Press 1969.

[9] V.D. Sánchez. *– Neurocomputing - Research and Applications -; – Increasing the Autonomy of Space Robots; – Intelligent BioSystems; – Modeling Dynamics for Communications, Navigation, Guidance and Control Applications*. German Research Center for Anthropotechnics, Wachtenberg-Werthoven, Germany; NASA Ames Research Center, Moffett Field, CA; KAIST Department of BioSystems, Daejeon, South Korea; Rockwell Collins, Advanced Technology Center (ATC), Cedar Rapids, IA, 1990, 2002, 2003, 2011.

[10] V.D. Sánchez. Neurocomputing *– Special Issue on Backpropagation, parts I-IV–, 5[4–6], 6[1–2]; – Special issue on RBF Networks, parts I-II –, 19[1–3], 20[1–3]; Advanced Support Vector Machines and Kernel Methods*, 55[1–2], 1993–1994, 1998, 2003.

[11] V.D. Sánchez et al. Maschinenmarkt *– On the Way to Intelligence, Structure and Function of Artificial Neural Networks using Supervised Learning; – The Grey Cells as Example, Analyzed Neural Operations can be realized by VLSI Components (in German) –*, 96[46], 97[3]; Chip Plus *– ANSpec, A Specification Language (in German) –*, 7; Technische Rundschau *– Neurocomputers in Industrial Applications (in German) –*, 82[65]; Neurocomputing *– The Design of a Real-Time Neurocomputer Based on RBF Networks –*, 20, 1990, 1991, 1990, 1990, 1998.

[12] V.D. Sánchez. *Modern Machine Learning Technology*. December 2017.
https://profdrvdsaphd.lima-city.de/documents/ModernMachineLearningTechnology.pdf

[13] V.D. Sánchez. *Continuous Big Data Applications in Industry – Modern Development Tools, Distributed Operational and Orchestration Systems, Internet of Things –*. December 2016.
https://profdrvdsaphd.lima-city.de/documents/ContinuousBigDataApplications.pdf

[14] V.D. Sánchez. *Computational Data Science Research and Technology Development – State of the Art –*. January 2023.
https://profdrvdsaphd.lima-city.de/documents/ComputationalDataScience.pdf

[15] V.D. Sánchez. *Deep Impact of Advanced AI Technology Developments – Government Regulatory Measures –*. September 2023.
https://profdrvdsaphd.lima-city.de/documents/ArtificialIntelligenceRegulation.pdf

[16] I. Goodfellow et al. *Deep Learning*. The MIT Press 2016.

[17] T.J. Sejnowski. *The Deep Learning Revolution*. The MIT Press 2018.

[18] M.P. Deisenroth et al. *Mathematics for Machine Learning*. Cambridge University Press 2020.

[19] G. James et al. *An Introduction to Statistical Learning – with Applications in R –*. 2nd Edition, Springer 2023.

[20] J. Roberts. *Gaining An Edge In Life & Business With AI: Unleashing the Power of Generative AI and Chat GPT*. Piper Publishing, 2023.

[21] E. Mollick et al. *Generative AI: The Insights You Need from Harvard Business Review*. to be released January 2024.

[22] S. Ozdemir. *Quick Start Guide to Large Language Models: Strategies and Best Practices for Using ChatGPT and Other LLMs*. Addison-Wesley 2023.

[23] M. Sewak et al. *Practical Convolutional Neural Networks : Implement advanced deep learning models using Python*. Packt 2018.

[24] K. Simonyan and A. Zisserman. *Very large convolutional networks for large-scale image recognition*. 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA.

[25] A. Krizhevsky et al. *ImageNet Classification with Deep Convolutional Neural Networks*. 26th Annual Conference on Neural Information Processing Systems (NIPS25, 2012), Lake Tahoe, NV, USA.

[26] P. Werbos. *Generalization of Backpropagation with Application to a Recurrent Gas Market Model*. Neural Networks 1, 339–356, 1988.

[27] F.M. Salem. *Recurrent Neural Networks – From Simple to Gated Architectures –*. Springer 2022.

[28] S. Hochreiter and J. Schmidhuber. *Long Short-Term Memory*. Neural Computation 9(8), 1735–1780, 1997.

[29] A. Vaswani et al. *Attention Is All You Need*. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

[30] Open AI. (2023). *GPT-4 Technical Report*. arXiv: 2303.08774v4 [cs.CL].

[31] S. Bubeck et al. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4*. arXiv: 2303.12712v5 [cs.CL].

[32] Nvidia. *NVIDIA DGX GH200 AI Supercomputer – AI Supercomputer for the Generative AI Era –*. White Paper, WP-11400-001 v01, June 2023.
https://resources.nvidia.com/en-us-dgx-gh200/technical-white-paper

[33] Nasscom, Deloitte. *Large language Models (LLMs) – A Backgrounder*. September 2023.

[34] V. Lialin et al. (2023). *Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning*. arXiv: 2303.15647v1 [cs.CL].

[35] A. Radford et al. (2023). *Language Models are Unsupervised Multitask Learners*. OpenAI, 2019.

[36] J. Pfeiffer et al. (2022). *Modular Deep Learning*. arXiv: 2302.11529v1 [cs.LG].

[37] J. Weizenbaum. *ELIZA – A Computer Program for the Study of Natural Language Communication Between Man and Machine*. Communications of the ACM 9(1), 36–45, January 1966.

[38] Department of Defense. *Data, Analytics, and Artificial Intelligence Adoption Strategy – Accelerating Decision Advantage –*, June 27, 2023.

[39] White House. *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. October 30, 2023.

[40] Congressioal Research Service. *Highlights of the 2023 Executive Order on Artificial Intelligence for Congress*. November 17, 2023.

[41] G. Miller. *US Senate AI 'Insight Forum' Tracker*. TechPlolicy.Press, December 8, 2023.
https://www.techpolicy.press/us-senate-ai-insight-forum-tracker/

[42] Special Competitive Studies Project (SCSP). *Generative AI: The Future of Innovation Power*.
https://www.scsp.ai/reports/gen-ai/