Artificial Intelligence Beyond Large Language Models

Colonel Ashish Nagaich, SM[@]

Dr Gopal Bhushan[#]

Abstract

This article examines the long-standing quest to make machines intelligent and emulate human behaviour. However, the explosive popularity of Large Language Models (LLMs) like ChatGPT, Llama, and Gemini, etc., since their launch in 2022 has not only accelerated the proliferation of Artificial Intelligence (AI) but has also created a misconception that AI is predominantly about these transformer-based LLMs. This myopic view ignores the multifaceted AI landscape, comprising of numerous models, which complement and outperform LLMs in specialised domains. This article looks at the journey of AI till date and explores its universe beyond LLMs, with an attempt to holistically understand the technologies and foster their innovation and exploitation in various fields. A clear understanding of the diverse requirements of these models—particularly in terms of computing and their applications—will enable practitioners and policymakers to responsibly leverage their full potential in addressing real-world challenges.

[@]**Colonel Ashish Nagaich, SM**, a 1999 batch Signals officer, is an alumnus of National Defence Academy and Indian Military Academy and holds an MTech in Computer Science and Information Technology. He has extensive operational experience across plains, mountains, deserts, and counter-insurgency areas. His staff appointments span operations, projects, training, and planning in India and abroad, including Sudan and Vietnam. He holds two master's degrees, is pursuing his PhD in Responsible Artificial Intelligence, and is presently working as a Senior Research Fellow at the United Service Institution of India.

[#]**Dr Gopal Bhushan** is an innovative professional with over 35 years of experience in technology management and corporate administration at Defence Research and Development Organisation. He has held senior roles including Directorial positions, Head of Defence Technology Wing (Indian Embassy at the United States), and global head of International Directorate. He currently serves as Deputy Director General at Amity Directorate of Science and Innovation, Noida.

Journal of the United Service Institution of India, Vol. CLVI, No. 643, January-March 2026.

Introduction

Artificial Intelligence (AI) is arguably the most transformative technology, which has started to have a profound impact on all possible fields of the world today. It encompasses "The study of agents that receive percept from the environment and perform actions", as defined by Russell and Norvig[1], and aims to replicate human intelligence in multiple domains. There has been a paradigm shift wherein machines, which have been performing their tasks based on deterministic logic till date, are operating on probabilistic models developed by AI. Though developments in all sub-fields and branches of AI have been significant and note-worthy, it is due to the human connect with AI systems through the Natural Language Processing (NLP) interfaces that Large Language Models (LLMs) have gained disproportionate popularity. The enhanced accuracy and exponential proliferation of deep neural networks based LLMs have created a general perception of AI being synonymous to LLMs. Tools like ChatGPT, Claude, and Gemini have not only excelled in understanding the context of conversations and queries but have become increasingly proficient in generating human-like texts, software codes, and even scientific research, thanks to the investments exceeding USD 100 bn in 2024 alone towards research and development in the field.[2] The versatility of LLMs make many believe them to be a one-point solution to all the AI problems, overlooking their limitations in domains beyond sequential data. LLMs only excel in NLP and falter in domains like causal reasoning, explainable rule-based decision making, novel designs, and long-term planning, especially in resource-constrained environments. The article traces back the historical development of AI and its basic components, with an aim to understand the genesis, mechanics, common applications, use-cases, and hardware requirements for various 'Types' of AI.

Journey Thus Far

The concept of intelligent machines was first conceived by Alan Turing[3], who proposed the 'Turing Test' to measure machine intelligence, but it formally got traction in the Dartmouth Conference[4], when John McCarthy and his colleagues worked on the idea. The 'Good Old-fashioned' AI mainly relied on symbolic logic and could prove mathematical theorems.[5] The first NLP system, called Eliza in 1966, was nothing but a set of pre-fed

answers to probable questions. A breakthrough came in the form of 'Recurrent Neural Networks' in 1972, which were designed to process sequential data like text, speech, and time series by using internal memory to retain information from previous inputs, saved in the form of 'State' of the system. However, the 1970s and 80s saw 'AI Winters' due to hype mismatch, predominantly due to a lack of corresponding development in the field of hardware, which made processing extremely slow and cumbersome. These systems had a potential of getting trained via 'Backpropagation Through Time'.[6] The supervised learning of AI models by human beings led to an evolution of numerous systems, for example, Apple's Siri in 2011 was trained to understand highly specific statements and commands, requiring human intervention. This evolved since the breakthrough development of artificial neural networks in 2012, which allowed machines to engage in reinforcement learning and simulate how the human brain processes information.[7] The processing time for these complex and heavy algorithms got significantly compressed by the 'Transformers' architecture, proposed by the Google and DeepMind team[8], which led OpenAI to develop GPT-1 in 2018. The evolution was powered by the development of Graphical Processing Units (GPUs), which allowed parallel processing, thereby, significantly increasing the processing speed. Unidirectional (only words preceding the word) understanding of text was considerably enhanced by bi-directional (both preceding and succeeding) encoder representations from transformers, released by Google in 2018, which allowed understanding of the context better. Beyond this point, LLM technology saw limited architectural breakthroughs, but scale increased dramatically. With the release of GPT-3 in 2020—featuring 175 billion parameters—performance improved significantly, and the world began to recognise the true potential of LLMs. However, parallel advancements in other paradigms like 'Reinforcement Learning', 'Convolutional Neural Networks' and 'Generative Adversarial Networks', have been equally significant and have resulted in the development of numerous effective AI systems. However, despite considerable investments in AI technology, over 70 per cent of all industrial AI deployments still use non-deep methods like rule-based systems.[9] In the next section, the article explores the world of AI by understanding the classification and details of different models, including and beyond LLMs, their comparison, and applications.

Classification of Artificial Intelligence: Based on Capabilities

Based on its capabilities or the range of tasks that can be performed, AI can be categorised into three basic categories, as shown in Figure 1. Out of these three, 'Narrow AI' (or Weak AI) has been realised, with extensive research being underway in the fields of 'Strong AI' and 'Artificial Super-Intelligence'.

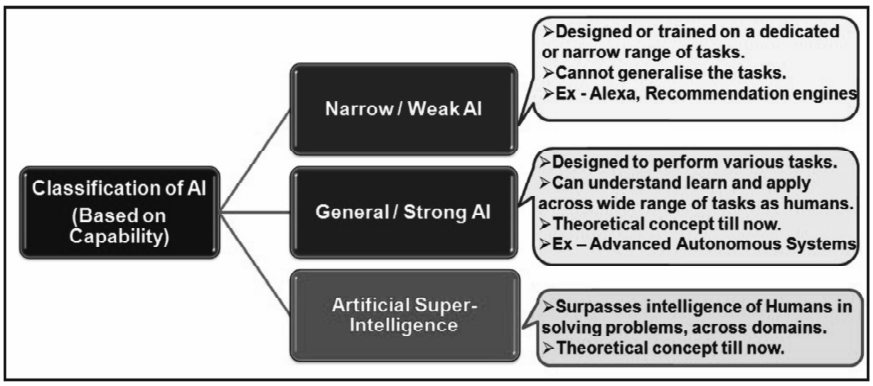

Figure 1: Classification of Artificial Intelligence based on Capability

Traditional Branches of Artificial Intelligence

Machine Learning (ML). It teaches computers to learn from data and become smarter over time. Supervised ML trains the system on data that has known answers or examples, like showing pictures of mountains to identify the same. Unsupervised ML, on the other hand, learns from data without knowing the answers ahead of time, by finding patterns or grouping them on their own. Reinforcement learning rewards for making good decisions and penalises for making a wrong one to improve accuracy. LLMs rely heavily on self-supervised pre-training. ML has been central to AI systems being used for demand forecasting, product recommendations, pattern detection, and predictive maintenance.

Speech Recognition. It helps in converting text to speech and speech to text and is useful in voice sample processing applications like interception of voice signals.

NLP. It deals with teaching computers to understand languages and interact in a way like humans do. It allows machines to read, write, speak, and respond sensibly. Applications include text analysis, translation and summarisation, grammar checking, and virtual assistants.

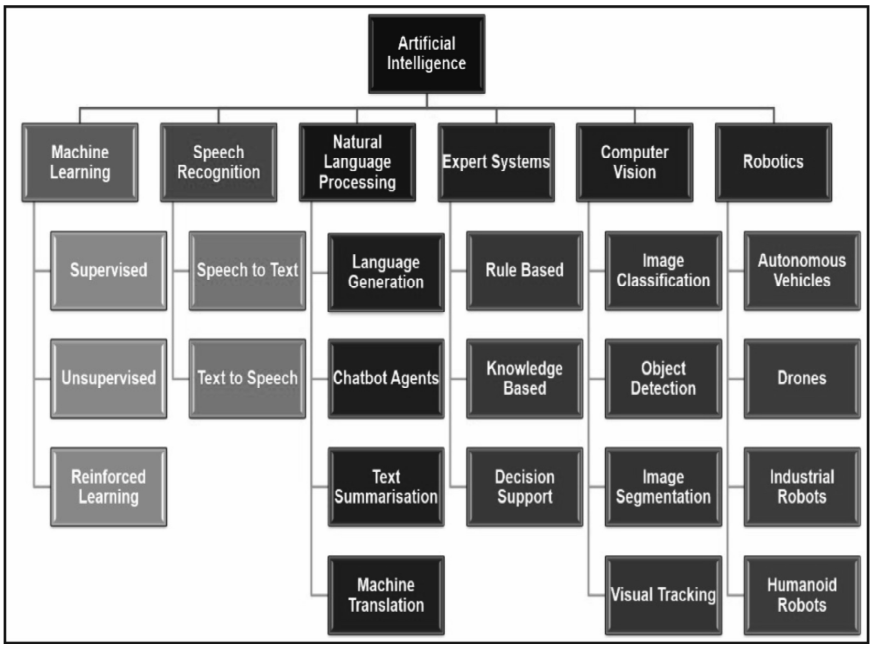

Figure 2: Traditional Important Branches of Artificial Intelligence

Expert Systems. They mimic human decision-making with use of rules and facts to solve specific problems like medical diagnostics, financial services, and transparent functioning in a regulated environment. 'Fuzzy Logic' deals with reasoning, helping systems to make decisions when the situation or information is uncertain or imprecise like control systems of appliances and embedded systems.

Computer Vision. It helps computers to understand and interpret visual information, including images and videos. Its application includes fields like image recognition, detection, multimodal vision, and image and video analysis for satellite, sensor, security cameras, and medical imaging.

Robotics. It deals with designing and programming robots to perform tasks, with or without human help. This rapidly growing field has many applications like factory automation, assistive robots for disabled and unmanned systems, including drones, etc. Multiple branches like computer vision, ML, and decision-making functions are combined to evolve 'Collaborative Robots' (Cobots) to work alongside humans.

Interplay between Branches. Human cognition is a complex function of all sensory inputs. Similarly, advanced AI systems are a combination of multiple branches, which support and complement each other, exploiting multiple technologies or concepts for improved efficiency, blurring the erstwhile distinction. For example, a system may combine vision and speech recognition for inputs, ML for prediction, NLP for user interface, and an expert system for rules and compliance. In practice, hybrid systems are so advanced that single-branch systems are seldom deployed. Blurred distinction between the branches has led to evolution of a more application-oriented model, based on the following two core technologies:

- Neural networks mimic functioning of human brain with interconnected nodes that process information and maintain states and memory. Deep neural networks-based learning uses multiple layers of neural networks to learn from large amounts of data and is used for applications like speech, text, and image recognition and processing.

- Generative AI can establish and learn patterns between inputs to create new content like text, images, music, or videos. These systems unlock exciting possibilities for creativity and innovation but are also laden with risks of deepfakes.

Functional Components of Artificial Intelligence. These fundamental components combine to give the desired functionality and accuracy to the AI systems. They offer modular capability to the complex AI systems and rely on the following underlying technologies.

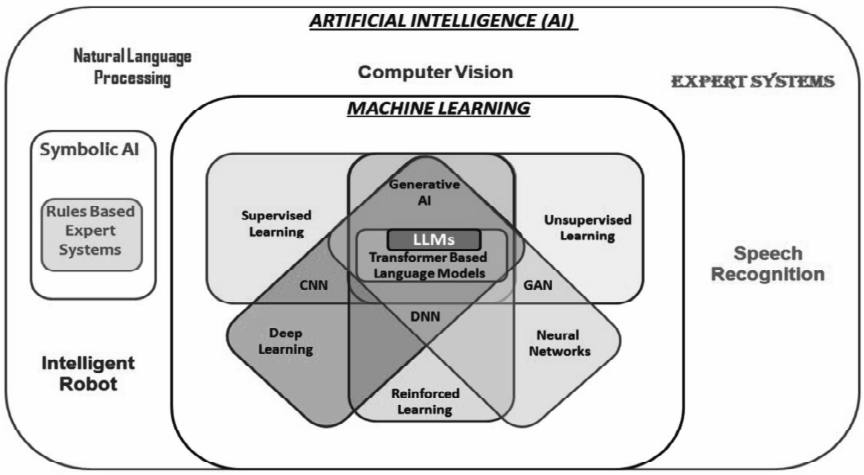

Figure 3: Interplay of Functionalities

- **Symbolic AI: Sticks to Rules**. Symbolic AI uses simple, explicit, human-readable rules and logic to infer, reason, and solve problems deterministically. It, therefore, offers a high degree of interpretability. However, since not all problems can be addressed using rigid logic such as if-then-else rules, this approach lacks flexibility. In contrast, LLMs are probabilistic, neural network-based systems that discover patterns from vast datasets to generate text, offering high flexibility but poor interpretability. Core techniques of symbolic AI include knowledge representation, automated theorem proving, and planning algorithms like Stanford Research Institute Problem Solver.[10] They require lesser training data than LLMs, which hallucinate.[11] Symbolic AI is, therefore, often used in a hybrid mode as it performs well in verifiable domains like rule-based problem solving, law, or aviation, for e.g., autonomous systems, checking legal contracts, and playing rule-based games.

- **Reinforcement Learning (RL): Learning from Interaction**. RL trains models from millions of interactions and tries to replicate mastering through penalties and rewards. This is like a child trying to learn cycling, where balance is a reward while falling can be considered as a penalty. The aim of the algorithm is to maximise the long-term wins.[12] RL in 'DeepMind's AlphaGo' beat World Chess players[13] by learning the game through replaying it multiple times.[14] Other

examples of RL in dynamic situations include self-driving cars, algorithms learning stock trading, or power grids learning to balance voltage fluctuations and solar spikes. This RL is different from text predictions by LLMs, which tries to match RL through human feedback[15] for fine tuning.

- **LLMs: Churn Text Based on Context**. LLMs are general-purpose foundational neural network models trained using deep learning on huge amount of text from books, websites, articles, academic papers, conversations, and code repositories. The text is broken down into 'Tokens', which are plotted on a multi-dimensional plane to understand their usage based on 'Statistical Pattern' of the language. Transformer neural networks enable parallel processing and contextual awareness to understand relationships between words and resolve ambiguity. This helps systems to predict, understand, generate, and reason in human language and perform multiple tasks like chatbots, summarisation, translation, analysis, and coding. The lack of additional hardware requirements, combined with a user-friendly interface, has helped democratise AI. As it learns the subject through a series of word-usage, it can invent facts, fabricate citations, and sound confident even when wrong, or 'Hallucinate' and, at times, lack grounded reality and have bias based on the training data.

- **Evolutionary Algorithms (EA): Evolves Like Nature**. EAs, also called 'Genetic Algorithms', mimic Darwinian evolution based on mutation, crossover, and selection of the fittest to optimise solutions. It uses trial and error method to handle non-differentiable, multimodal problems. The system selects a random solution to check its correctness and accuracy. If not found satisfactory, it keeps improving it by adaptation and amendments till the desired level of the perfection is reached. These algorithms are, therefore, fundamental in deriving new concepts and designs, for e.g., shape and design of National Aeronautics and Space Administration's Mars ST5 antenna was evolved to outperform human engineers, reducing weight by 44 per cent while boosting gain.[16] Other applications include route optimisation by airlines and taxis to save fuel games, where AI opponents evolve their tactics and adapt mid-match. An associated, more

advanced, and specialised subset of EAs is neuro-evolution, which applies evolutionary principles to artificial neural networks, optimising their weights and topology to create intelligent and functional neural networks. A few examples of neuro-evolution include OpenAI's early agents and complex climate modelling systems.

- **Probabilistic Models: Uncertainty and Causality**. Probabilistic Graphical Models (PGMs) capture dependencies and causal relationships, using Bayesian probability principles to infer outcomes.[17] They quantify uncertainty and update beliefs with evidence or causes. A graph of influences or dependent factors is drawn and a 'What If?' question is asked to arrive at a reasonable answer.[18] PGMs use causality to find answers or reasons, which are critical for scientific applications including prediction of ailments from symptoms. They are also used in spam filters and user-specific recommendations for advertisements for Microsoft[19], which increased their returns by 20 per cent and over-the-top platforms. These models differ from LLMs, which predict succeeding words probabilistically, but do not consider the cause or rationale for the same.

- **Neuro-Symbolic Models: Brains with Logic**. Neuro-symbolic AI models use a hybrid approach by doing pattern-matching using neural networks and logic reasoning, thereby, achieving the best of both worlds. Examples include 'Alpha Geometry', where LLM guesses the steps and logic of symbolic models and verify them to solve geometric problems.[20] AI systems like neural theorem provers logically prove the conjectures with 60 per cent accuracy.[21] Various limitations of LLMs like biases and hallucinations can be mitigated using symbolic verification by neuro-symbolic models, thereby, enhancing their accuracy.

The Future—Mixing Models for Smarter Artificial Intelligence

The future of AI will not be defined by single models, but coordinated operations by multiple specialist models, each performing their specific tasks optimally and mitigating each other's shortcomings. For example, robots like Google's RT-1 grabs objects using RL with vision.[22] Though instrumental in deep proliferation of AI, LLMs suffer from numerous limitations, which can be

overcome by other models, as shown in Table 1, to develop much more complex and efficient systems.

LLM: Limitations	Models which Mitigate	Strength of Models	Applications
Hallucinations	Symbolic AI	Explainable rules	Legal contracts
No true invention	Evolutionary AI	Novel designs	Antenna design
No real agency	RL	Sequential decisions	Game AI, Robots
Ignores causality	Probabilistic Models	Handles uncertainty	Medical diagnosis
Poor logic	Neuro-Symbolic AI	Reasoning and patterns	Math Proofs

Table 1: Limitations of Large Language Models and Complementing Models

Model-Specific Requirements

Each of the above-mentioned models process the data differently and, therefore, require different types of hardware to perform optimally. Processors and memory required for running these complex models efficiently vary and so does their cost and deployment model. The way models process data dictates the requirement of parallel processing and memory, which is met by GPUs and onboard memory. For example, LLMs with billions or trillions of parameters require GPUs for large-scale parallel processing and specialised, high-speed memory to store graphics data like frame buffers and 3D models called Video Random Access Memory (VRAM). On the contrary, symbolic AI and EAs thrive on low-end Central Processing Units (CPUs) or Field Programmable Gate Arrays (FPGAs). Similarly, symbolic or probabilistic models are light on computation, allowing functioning on edge devices. RL and LLMs have higher compute requirements, which necessitate working through GPUs and storage hosted in data centres. Typical model-specific requirements of hardware to include compute and memory are summarised in Table 2, a thorough understanding of which will facilitate decision and policy makers in selecting the correct model, architecture, and cater for the infrastructure requirements for fielding of AI systems.

Paradigm or Model	Typical Hardware	VRAM Requirement	Compute Requirement	Typical Cost or Power
LLMs (e.g., GPT-4 or Llama)	High-end GPUs (A100/H100), 24 to 48 GB+ VRAM per model	Extreme (32B+ models: 48GB+)	GPUs or Tensor Processing Unit (TPUs) for training or inference	User Data Device (UDD) 10,000; high power
Symbolic AI (PROLOG)	Standard CPU, embedded chips (1980s: custom AI boards ~600 KLIPS)	Low (MBs)	CPUs, FPGAs for speed	Laptop (UDD 500), very low power
Evolutionary Algos (GAs)	CPUs, FPGAs for parallel evals; ScienceDirect+1	Low-medium (GBs for populations)	Parallel CPUs or FPGAs	Mid-range FPGA (UDD 1,000), efficient
RL (AlphaGo or MuZero)	TPUs or GPUs (MuZero: 16-1000 TPUs)	High (tens GBs)	GPU or TPU clusters	Cloud clusters (UDD 100,000 equivalent), high power
Probabilistic (Bayesian Nets)	Central Processing Units (CPUs), brain-inspired hardware for efficiency	Low (for inference)	CPUs or custom Application-Specific Integrated Circuits	Standard server (UDD 2000), low power
Neuro-Symbolic (AlphaGeometry)	GPUs but memory-bound, less than pure deep learning	Medium-high (varies by logic operations)	GPUs with optimisations	Mid-high GPU (UDD 5000), Moderate power
Multimodal (Contrastive Language–Image Pretraining or Neural Radiance Fields)	Multi-GPU (A100/RTX), high-core CPUs	High (256GB+ Random Access Memory for embeddings)	GPUs + fast storage or Network Interface Cards	Hybrid rig (UDD 20,000), high power

Table 2: Hardware Requirements of Different Models

Technology Prognosis

The direction of research and evolution of AI suggests that hybrid and multimodal AI systems will define the future. They will utilise neural networks for learning and analysing data sets, symbolic AI for rule- or facts-based logical reasoning, and probabilistic models to take decisions in uncertain scenarios. These multiple models, employed in various combinations, will not only mitigate the shortcomings of each other but would also enhance accuracy and alignment of the system output, making them more explainable. Agentic AI, which uses interactions between different models, is a natural progression in this direction, and collaborative agents are predicted to be the next big thing in the realm of AI. Multiple AI agents working together on multifarious subjects could collaborate in real time to give cross-domain intelligence, contextual understanding, and give rise to strong AI or artificial general intelligence.

Conclusion

An understanding of various AI models and their underlying technology is imperative for decision making and policy formulation towards deployment of AI systems. It is recommended that suitable combination of AI models be identified as the first step towards planning and fielding of these systems to accomplish the requisite functions or applications. This will dictate the exact requirement of different types of hardware and their hosting architecture. A thorough understanding of functioning and limitations of these models will not only streamline the process of fielding of AI systems but will also make their exploitation more effective. For example, a sizable share of common applications could run on small language models utilising commonly available and cheap CPUs instead of costly GPUs. Thus, fielding of AI systems could be optimised and expedited in a cost-effective manner. This would immensely help countries like India, which does not yet have the requisite fabrication facility and is entirely dependent on imports for advanced information technology hardware.

Endnotes

[1]Stephen Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. (Pearson: Harlow, 2021).

[2]McKinsey & Company, *State of AI Report* (McKinsey & Company: New York, 2025).

[3]AM Turing, "Computing Machinery and Intelligence", *Mind,* 1950, accessed on 12 Feb 2026, https://courses.cs.umbc.edu/471/papers/turing.pdf

[4]John McCarthy et al., "Dartmouth Summer Research Project on Artificial Intelligence", *Dartmouth College,* 1956, accessed 15 Feb 2025, http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf

[5]Allen Newell and Herbert A. Simon, "The Logic Theory Machine—A Complex Information Processing System", *IRE Transactions on Information Theory* 2, no. 3, Sep 1956, accessed 16 Feb 2026, DOI: 10.1109/TIT.1956.1056797

[6]David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams, "Learning Representations by Back-Propagating Errors", Nature, 323, no. 6088, p 533, 1986, accessed 18 Feb 2026, https://doi.org/10.1038/323533a0.

[7]Ankur Jain, “Demystifying NLP & LLM”, *Website Page of Ankur Jain*, 24 Sep 2023, accessed 20 Feb 2026, https://iankur.com/artificial-intelligence/demystifying-nlp-llm/

[8]Ashish Vaswani et al., “Attention Is All You Need”, in *Advances in Neural Information Processing Systems 30*, ed. I. Guyon et al., (NeurIPS: Denver, 2017), pp 5998–6008, accessed 22 Feb 2026, https://papers.nips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[9]Gartner Insights Abstract, “AI Adoption Trends: Leaders Define 3 Outcomes Across Verticals”, *Gartner,* 2025, accessed 25 Feb 2026, https://www.gartner.com/en/documents/7206130#:~:text=Summary,dominate%20a%20$3.3%20trillion%20market

[10]Richard E Fikes and Nils J Nilsson, “STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving”, *Artificial Intelligence,* 2, no. 3–4, pp 189–208, 1971, accessed 26 Feb 2026, https://www.scirp.org/reference/referencespapers?referenceid=62039

[11]Ziwei Ji et al., “Survey of Hallucination in Natural Language Generation”, *ACM Computing Surveys,* 55, no. 12, Mar 2023, 27 Feb 2026, https://doi.org/10.1145/3571730

[12]Richard Sutton and Andrew Barto, Reinforcement Learning: An Introduction, 2nd edition, (Cambridge, MA: MIT Press, 2018).

[13]David Silver et al., “Mastering the Game of Go with Deep Neural Networks and Tree Search”, *Nature,* 529, no. 7587, pp 484–489, 28 Jan 2016, accessed 25 Feb 2026, https://doi.org/10.1038/nature16961

[14]Julian Schrittwieser et al. “Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model”, *Nature,* 588, no. 7839, pp 604–609, Dec 2020, accessed 28 Feb 2026, https://pubmed.ncbi.nlm.nih.gov/33361790/

[15]Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, et al, “Training Language Models to Follow Instructions with Human Feedback”, in *Advances in Neural Information Processing Systems, (*NeurIPS: New Orleans, 2022), eds by S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, and A Oh, pp 27730–27744, https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html

[16]Jason D Lohn, Gregory S Hornby, and Derek S Linden, “An Evolved Antenna for Deployment on NASA’s Space Technology 5 Mission”, in *Genetic Programming Theory and Practice II*, eds. Una-May O’Reilly et al. (Boston: Springer, 2005), pp 01-10.

[17]Daphne Koller and Nir Friedman, *Probabilistic Graphical Models: Principles and Techniques* (Cambridge, MA: MIT Press, 2009).

[18]Judea Pearl, Causality: Models, Reasoning, and Inference, 2nd ed. (Cambridge: Cambridge University Press, 2009), pp 100-105.

[19]"CausalML Documentation", *Microsoft*, 2023, accessed 25 Feb 2026, causalml.readthedocs.io

[20]Trieu H Trinh et al., "Solving Olympiad Geometry Without Human Demonstrations", *Nature* 625, no. 7995, p 478, 17 Jan 2024, accessed 28 Feb 2026, https://doi.org/10.1038/s41586-023-06747-5

[21]Balaji Rao, William Eiers, and Carlo Lipizzi, "Neural Theorem Proving: Generating and Structuring Proofs for Formal Verification", *arXiv preprint*, 2025, accessed 08 Mar 2026, https://arxiv.org/abs/2504.17017v1

[22]Anthony Brohan et al., "RT-1: Robotics Transformer for Real-World Control at Scale," *arXiv preprint, 2022,* accessed 23 Feb 2026, arXiv:2212.06817 (2022), https://doi.org/10.48550/arXiv.2212.06817