

Artificial Intelligence and the Evolution of AI (Model) Capabilities

A Conceptual Primer



Published in 2025 by the United Nations Educational, Scientific and Cultural Organization
7, place de Fontenoy, 75352 Paris 07 SP, France

© UNESCO 2025

ISBN 978-92-3-100737-8

DOI: <https://doi.org/10.54678/DDEC6804>



This publication is available in Open Access under the Attribution-ShareAlike 3.0 IGO (CC-BY-SA 3.0 IGO) license (<http://creativecommons.org/licenses/by-sa/3.0/igo/>). By using the content of this publication, the users accept to be bound by the terms of use of the UNESCO Open Access Repository (<https://www.unesco.org/en/open-access/cc-sa>).

Images marked with an asterisk (*) do not fall under the CC-BY-SA license and may not be used or reproduced without the prior permission of the copyright holders.

The designations employed and the presentation of material throughout this publication do not imply the expression of any opinion whatsoever on the part of UNESCO concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

The ideas and opinions expressed in this publication are those of the authors; they are not necessarily those of UNESCO and do not commit the Organization.

Authors: Roman Jurowetzki (Aalborg University) and Mariagrazia Squicciarini (UNESCO)

Cover photo: : © Tapati Rinchumrus/Shutterstock.com*

Cover illustration and Graphic Design: UNESCO/Jzan Tamielle Villanos

Printed by UNESCO

Printed in France



SHORT SUMMARY

Understanding AI Capabilities in 2025

With artificial intelligence (AI) making remarkable advances, understanding AI's key components, modus operandi and capabilities becomes paramount for societal and policy purposes, to be able to assess possibilities and challenges beyond what is claimed for marketing or lobbying purposes.

This UNESCO primer provides essential technical insights for decision-makers and citizens alike, through an analysis of AI's core architecture, capabilities and developments.

Over the 2023-24 AI-related deployment and adoption possibilities as well as economic prospects evolved dramatically: While development costs remained high, usage costs for frontier models dropped by over 90% due to efficiency improvements and increased competition, significantly facilitating and accelerating access to AI applications and advancements. While this can help improve people's lives, it may also lead to harmful applications of powerful technology if not ethically guided.

90%

Drop in AI processing costs can enable wider access and greater adoption, but triggers ethical challenges

Leading AI systems now match human expert performance in specific domains, though benchmark results require careful interpretation for real-world applications. Open-source alternatives are rivaling proprietary systems, lowering innovation barriers while raising misuse concerns.

This evolution demands adaptive governance frameworks that balance oversight with innovation potential, calling for dynamic policy approaches that can evolve alongside AI capabilities.



"Since wars begin in the minds of men and women it is in the minds of men and women that the defences of peace must be constructed"

Artificial Intelligence and the Evolution of AI (Model) Capabilities

A Conceptual Primer

Table of Contents

| | |
|---|-----------|
| Foreword | 3 |
| Acknowledgements | 4 |
| Introduction | 5 |
| Artificial Intelligence’s Key Concepts and Technological Trajectories | 7 |
| Key technological trajectories in AI | 7 |
| Modalities of Learning in AI | 9 |
| Fundamental Concepts | 10 |
| Tracing the Evolution of AI (Model) Capabilities | 12 |
| Traditional Machine Learning Capabilities | 12 |
| Deep Learning Breakthroughs | 14 |
| Generative Abilities and Multimodal Integration | 15 |
| Code Generation and Automation | 17 |
| Reasoning, AI Agents, and Autonomy | 19 |
| Model Evaluation | 21 |
| The Hardware Layer | 22 |
| AI Systems, Designs, and User Interfaces | 23 |
| The Co-Evolution of AI Capabilities and Governance Challenges: Some First Conclusions | 24 |
| Glossary | 25 |
| References | 29 |

Foreword



© UNESCO/ Christelle Alix

Not one single day goes by without very big news about significant investment, advancement and applications in Artificial Intelligence (AI). Examples are DeepSeek's release of R1 open-source model on 20 January 2025, which matches or exceeds the capabilities of leading proprietary systems while costing just a fraction of that. R1 came out only a few days after UC Berkeley researchers announced that they had replicated OpenAI o1- preview's performance. On 21 January 2025 the US Government announced up to 500 billion dollars in private sector investment to build artificial intelligence infrastructure in the United States; while OpenAI, Softbank and Oracle disclosed plans for an AI-related joint venture called Stargate.

This hype around AI is mirrored in citizens' and policy makers' conversations, which recurrently express their fears and hopes about something that, nevertheless, few know well. Too often, many talk about AI without having a clear understanding about what AI is and is not, what it can do or is unable to, and how AI systems actually work — their capabilities, limitations, and the ways they learn and make decisions. Lacking such fundamental knowledge, the risk is that discussions may become polarized and policy design (at best) inaccurate, and even end up either stifling innovation or failing to adequately protect against genuine risks. A clear understanding of the specific technologies, and how they have evolved from more basic rule-based systems to generative AI and increasingly complex agentic systems is paramount for the design and implementation of effective and adaptive governance frameworks, as those advocated in UNESCO Recommendation on the Ethics of Artificial Intelligence.

This primer aims to fill this gap. It offers a comprehensive yet accessible examination of AI's fundamental building blocks, its evolving capabilities, and the complex interplay between technical advancement and governance challenges. By providing this foundation, it aims to enable a more informed dialogue between technologists, policymakers, business leaders, and society at large. At stake is not just technological progress, but the future of human agency and decision-making in an AI-enabled world, our human rights, our fundamental freedoms.

Gabriela Ramos

Assistant Director-General for Social
and Human Sciences, UNESCO

Acknowledgements

This report was written by Roman Jurowetzki (Aalborg University) and Mariagrazia Squicciarini (UNESCO) under the leadership and overall supervision of Gabriela Ramos, Assistant Director-General for Social and Human Sciences at UNESCO.

Our sincere gratitude goes to Syed Quiser Ahmed (Infosys) for providing very helpful comments, as well as to Max Kendrick; to Jzan Tamielle Villanos for editing and beautifying the paper; and to Salma Ibrahim and Jose Valdez Genao for proofreading the paper.

All errors remain our own.

Introduction

In recent years, it has become evident to users and policy makers alike that the development, deployment and use of Artificial Intelligence (AI) needs being steered by guardrails ensuring that AI's transformational potential upholds human rights, fundamental freedoms and the rule of law, for the good of economies and societies.

As a point in case, the 2024 G20 Declaration of Ministers responsible for the Digital Economy states: "We recognize that safe, secure and trustworthy Artificial Intelligence (AI), when applied in a transparent, ethical, responsible and reliable manner, may act as a catalyst for achieving economic growth and inclusive sustainable development within its three dimensions: social, economic and environmental. We reaffirm the G20 AI principles and the UNESCO Recommendation on the Ethics of AI" and "...our commitment to leverage AI for good and for all, as well as to unlock the full potential of AI, share its benefits for all and mitigate its risks."¹

On the one hand, to design and implement tools able to effectively pursue these objectives, policy makers need having a clear understanding of what AI technologies are, what they can and cannot do, how they function, and how they evolve. In the absence of such broad knowledge, the risk is that policy may end exacerbating the very problems it is trying to fix in the first place. On the other hand, individuals and societies need being aware of what this set of transformational technologies are and what they may bring to them, in terms of both risks and opportunities, to be able to take informed decisions about interacting with and leveraging AI.

These concerns intensify as AI capabilities expand, and rapidly so, and more sophisticated systems emerge. As AI progresses from automation to complex cognitive tasks, so do concerns related to possible the acceleration or broadening of job displacements, given AI's ability to perform a wide range of cognitive and creative tasks. Heightened privacy concerns emerge as multimodal AI systems can process and correlate diverse types of personal data, exposing individuals to abuses and discrimination, and impinging upon their independence and agency, among others. Generative AI continues to challenge information authenticity and integrity, through the creation of increasingly convincing synthetic content made of text, images, audio and videos. The ethical implications of AI systems featuring advanced reasoning capabilities and autonomous decision-making abilities are not sufficiently discussed nor addressed, as are the consolidation of technological power and the massive computational resources that training advanced AI models requires. Safety concerns continue to be triggered as AI systems become more capable of independent planning and action through agentic capabilities. The environmental impact of AI technologies escalates as models grow larger and more computationally intensive. While many models are simultaneously becoming more efficient and compact, overall energy consumption continues to rise.

These developments amplify fundamental questions about preserving meaningful human agency and oversight as AI systems become more integrated into critical societal functions and decision-making processes become more compelling. These concerns are particularly pressing for individuals and policy makers' alike, especially as the gap between AI's capabilities and governance frameworks continues to widen.

The present primer, representing the state of AI understanding as of January 2025, aims to address this need for a broad understanding of what AI technologies are and what they do. Its goal is to represent a knowledge resource that, moving beyond a simple cataloging of AI technologies and applications, fosters a deeper understanding of how AI systems are constructed. To this end, it sheds light on the fundamental building blocks of such technologies and how they interact to create diverse functions and products that impact society, businesses, and every day's life.

The effective governance of AI-enabled applications requires the ability to identify and address the specific elements within these systems and the interfaces between them. These elements and interfaces define how tasks are solved and are often where system outcomes may be suboptimal or problematic. Also, by providing insights into how AI engineers approach and understand these technical building blocks, emerging capabilities, and their interactions, we hope to build bridges that facilitate a more nuanced and effective dialogue between technology practitioners and the broader public.

¹ <https://g7g20-documents.org/database/document/2024-g20-brazil-sherpa-track-digital-economy-ministers-ministers-language-g20-dewg-maceio-ministerial-declaration#section-2>

In what follows, we first discuss the key technological trajectories that have shaped AI developments, from rule-based systems to modern generative AI, highlighting their distinctive characteristics and implications. We then shed light on the various modalities of learning in AI, examining how different learning approaches—from supervised learning to zero-shot capabilities—have expanded AI’s adaptability and applicability. Following this, we explore the fundamental concepts underlying AI systems, including the crucial roles of data, training, and inference in shaping AI capabilities. We then trace the evolution of AI model capabilities across multiple dimensions, from traditional machine learning through deep learning breakthroughs, to the emergence of generative abilities and multimodal integration. This is complemented by an examination of recent developments in code generation, automation, and the emergence of AI models and agents with reasoning capabilities. In the final sections, we examine the practical aspects of implementing AI systems. We begin by exploring the challenges and methodologies of model evaluation, which has become increasingly complex following the emergence of generative AI systems. We then turn to the evolution of hardware infrastructure, tracing how computational requirements have shaped AI development and deployment. Finally, we analyze the critical role of system design and user interfaces in translating AI capabilities into practical applications, highlighting how these elements mediate between AI models and end users.

Artificial Intelligence's Key Concepts and Technological Trajectories

Apprehending the complexities of the rapidly evolving Artificial Intelligence (AI) landscape requires having some basic knowledge about key concepts and features characterizing the development and deployment of AI technologies, and the way these may facilitate or at times hinder progress. For instance, on the one hand, recent advancements in generative AI applications stem from specific developments in neural network architectures and training methodologies. On the other hand, challenges in AI adoption across industries frequently relate to constraints in data quality, model interpretability, or the need for specialized expertise. Such examples point to the fact that the variety of technological trajectories and the very pace at which AI developments are unfolding - from traditional rule-based systems to contemporary generative models – can have first order magnitude effects over industry penetration, technological adoption, market dynamics and, more generally, on what it may mean to live, work and produce in the AI era, and on what the AI future itself may look like.

Understanding these foundational elements further enables a more accurate assessment of the real capabilities and limitations of AI technologies. It can also help recognize the technical features and components that are more likely to influence market dynamics; infer whether AI technologies may complement or otherwise substitute roles, tasks and functions; and more accurately anticipate the potential directions that future AI developments may take. In what follows, we provide a succinct overview of the key AI concepts and mechanisms that are relevant to the analysis proposed in the present report, and pay attention to some of the pivotal developments that have shaped current AI's capabilities. Despite our best efforts to establish a clear structure that moves from abstract concepts to more concrete technological trajectories, some repetitions remain unavoidable, due to the overlapping and interdependent nature of trajectories and concepts. A case in point is transfer learning, which can be considered both as a technological trajectory and as a learning modality.

Key technological trajectories in AI

Table 1 compares the major AI technological trajectories emerged over the last years, and highlights their key characteristics and implications. Trajectories are presented in a way that mirrors historical developments and the way technological evolution has shaped modern AI capabilities, as well as their increasing complexity (Russell and Norvig, 2016; Goodfellow, 2016).

Rule-Based Systems represent the earliest form of AI. These systems employ predefined rules and logic to make decisions, rendering them effective for well-defined problems, but falling somewhat short in terms of flexibility and learning capabilities. The deterministic nature of these systems allows for explicitly described, transparent, and auditable actions. The creation of such systems relies on specialists defining the rules that the system must follow in different contexts, which limits their adaptability to novel situations. Common applications include traffic light control systems and traditional “smart” software.

Machine Learning (ML) marked a significant advancement in AI. ML systems are trained by data scientists or ML engineers to learn from data without being explicitly (re)programmed, improving their performance over time through exposure to data. However, as data must adhere to specific formats, the outputs generated by such systems exhibit limited complexity. ML is suitable for tasks like categorizing data (e.g., identifying potentially fraudulent credit card transactions or detecting spam) or predicting continuous values (e.g., estimating potential car prices). It is also widely used in recommendation systems (for content or products). Deep Learning, a subset of ML, employs artificial neural networks with multiple layers and excels at tasks such as image and speech recognition, allowing for more complex input processing. Despite improvements throughout the 2010s, ML requires expert knowledge (data scientists, engineers for data processing and deployment) and very specific data to be trained upon. Models are typically specialized and trained for particular use cases, with limited transferability to other applications.

Transfer Learning and Generative AI represent the latest frontier in AI development. Transfer Learning leverages pre-trained models and adapts their knowledge to new, related tasks. This approach significantly reduces the need for task-specific training data and computational resources (Pan and Yang, 2009). In contrast, Generative AI goes beyond adaptation to create new content (text, images, audio) that was not explicitly programmed. Large Language Models (LLMs) or more broader foundation models are a prime example of this set of technologies that excel at few-shot and zero-shot learning, facilitated by their size and emerging

Table 1: Comparison of AI Trajectories: Rule-Based Systems, Machine Learning, and Generative AI

| ASPECT | RULE-BASED SYSTEMS | MACHINE LEARNING | GENERATIVE AI |
|------------------|--|---|---|
| Input | Simple, structured data; well-defined parameters | Structured data (e.g., tabular, images); specific formats | Unstructured data (text, images, audio); high complexity |
| Processing | Explicit rules; deterministic logic; no learning | Statistical pattern recognition; task-specific training | Few-shot learning; emergent reasoning; context understanding |
| Output | Predetermined; limited complexity; predictable | Predictions and classifications; moderate complexity | Creative and diverse; high complexity; novel content |
| Adaptability | Low flexibility; manual updates; domain-specific | Moderate within trained domain; retraining for new tasks | High flexibility across domains; adaptable without retraining |
| Dev. Costs | Low initial; ongoing maintenance for rule updates | High: data collection, labeling, computational resources | Very high: initial R&D, training, ongoing research |
| User Costs | Minimal operation; potential customization fees | Data prep, model fine-tuning, inference computation | API fees or high-end hardware; potential fine-tuning |
| Expertise | Domain experts; software engineers | Data scientists; ML engineers; domain experts | Prompt engineers; domain experts |
| Scalability | Limited; linear growth with rule complexity | Good with data and compute; model-specific limits | Highly scalable; limited by compute resources |
| Interpretability | High transparency; easy to audit and explain | Moderate to low; depends on model complexity | Low transparency; challenging to interpret reasoning |
| Applications | Expert systems; decision trees; industrial control | Image classification; fraud detection; recommendations | Chatbots; content creation; code generation; emergent agentic systems |

Source: Authors' own compilation



reasoning properties (Brown et al., 2020; Radford et al., 2019). In late 2024, these reasoning abilities were more explicitly trained into the new generations of LLMs, including OpenAI's o1 and o3, Gemini Flash 2.0 Thinking, and DeepSeek R1, significantly enhancing their performance on complex tasks requiring logical reasoning, such as mathematics and science.

Few-shot learning allows AI models to make accurate predictions by conditioning them on a very small number of labeled examples, typically between 2 to 5 instances and is typically used to train models for classification tasks when suitable training data are scarce. Zero-shot learning enables models to perform tasks they were not explicitly trained on in the first place by leveraging auxiliary information or understanding relationships between different categories or labels, for example, recognizing different types of animals even if the model was only trained on a few specific species. LLMs are built to accept unstructured inputs, thus reducing the importance of data quality compared to traditional ML and allowing for application in diverse contexts. Training is typically performed by specialized AI labs, i.e. generally large companies having the capacity for resource intensive training runs. From a user perspective, the final model does not require training and makes on-the-fly adjustments based on input data and inferred tasks from instructions.

The evolution of AI from rule-based systems to machine learning and now to generative AI completely changes AI-related accessibility and application. Early AI required extensive in-house development, specialized data infrastructure, and considerable expertise, thus limiting its adoption primarily to well-resourced institutions and organizations. The move towards pre-trained models, particularly LLMs, now offers AI capabilities that are increasingly accessible through commercial services, reducing the need for local infrastructure and expertise. Such shift presents both opportunities and challenges for those applying AI: at the same time, while it may accelerate technology adoption by lowering barriers to entry for many applications, reliance on external services may exacerbate concentration, existing dependencies, and potentially lead to jeopardizing locally developed expertise and solutions. This points to the need to ensure equitable access and capacity building a way that accounts for local needs and ownership in the global AI ecosystem.

Modalities of Learning in AI

Learning modalities employed in Artificial Intelligence vary, with advanced techniques that enhance model flexibility and efficiency. These modalities form the foundation of many modern AI applications, and their capabilities dictate the approaches used in model training and deployment.

Supervised Learning represents a foundational approach in AI training. Here, models are trained on large datasets consisting of input-output pairs, where each input is paired with the correct output. The model learns to map inputs to outputs, gradually improving its performance through multiple iterations over the training data. This method requires extensive labeled datasets and is commonly used in tasks such as image classification or sentiment analysis (Caruana and Niculescu-Mizil, 2006). While effective, this approach is constrained by the need for large amounts of labeled data, which in turn often requires significant human effort in data preparation and annotation. The efficiency of the model is highly dependent on the quality and representativeness of the data used for training.

Transfer Learning marks an advancement in AI training and is a modality that allows models to apply knowledge gained from one task to a different albeit related task. For example, a model trained on general image recognition can be fine-tuned in order to perform a specific task like identifying medical conditions in X-rays. Transfer learning greatly reduces the amount of task-specific data needed and accelerates the training process. This approach has become increasingly popular, especially in domains where large labeled datasets are in short supply or expensive to obtain. As a result, transfer learning has effectively lowered the barrier for development in resource-constrained environments.

Few-Shot Learning addresses scenarios where limited labeled data are available for a new task. In this modality, models learn to generalize from just a few examples, typically between two to five instances per class, an approach that is particularly useful in real-world applications where obtaining large labeled datasets is impractical or prohibitively expensive. Few-shot learning often leverages meta-learning techniques, whereby the model learns to learn from limited data, enabling rapid adaptation to new tasks. This capability is particularly valuable to rapidly deploy AI models in diverse and changing real-world scenarios.

Zero-Shot Learning further expands the capabilities of AI models by enabling them to perform tasks they were not explicitly trained on. In this paradigm, models make predictions about classes they have never encountered during the training phase. This is achieved by using auxiliary information or by interpreting relationships between different classes. For instance, a model trained to recognize animals could identify a new species based on a textual description, without prior exposure to images of that very



species. While this capability significantly enhances the adaptability and generalizability of AI systems, it is important to note that the accuracy of zero-shot learning can vary substantially and often requires careful evaluation to ensure robustness and reliability.

One-shot learning bridges the gap between few-shot and zero-shot learning, with the model learning from only one example per new class. This modality is crucial in applications like facial recognition systems, where registering a new user with just one photo is desirable. One-shot learning often employs sophisticated techniques to extract maximum information from a single example, balancing generalization with specificity². However, one-shot learning can sometimes be less reliable than few-shot learning, as the model does not have enough exposure to learn from robust generalizations. With the rise of LLMs and other forms of generative AI, advanced learning modalities such as transfer, few-shot, one-shot, and zero-shot learning, are increasingly important (Wang et al., 2020; Xian et al., 2018). They enable AI systems to adapt more flexibly to new tasks and domains, reducing the reliance on extensive task-specific datasets and expanding the potential applications of AI across various fields. This trend is also lowering the barrier to AI adoption for many organizations: as long as a task is relatively general or easily codified, it makes it easier to use AI for automation purposes without developing and deploying a custom system. The shift towards these modalities allows for faster deployment, more efficient use of resources, and wider applicability, which are all key elements that can contribute to democratize access to AI technologies.

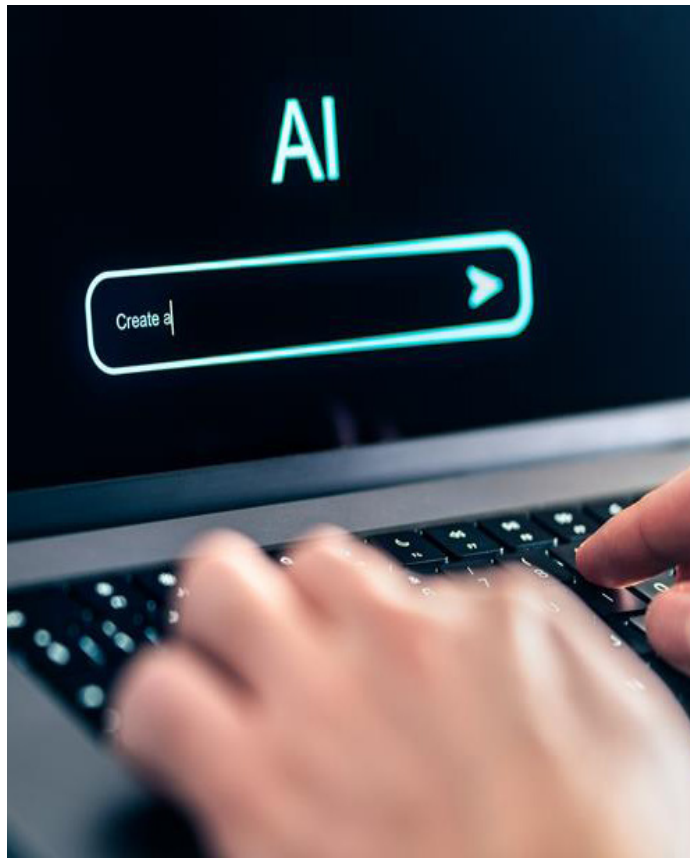


Image Source: © Tero Versalainen/Shutterstock.com*

The discourse around “prompt engineering” can be seen in light of these developments toward more advanced learning modalities. A prompt is text describing the task that AI should perform and, consequently, prompt engineering refers to the process of forming instructions that can be interpreted and performed by e.g. generative AI models (Liu et al., 2023). This represents a fundamental shift in how AI solutions are developed and deployed, potentially reducing the technical barriers to AI adoption while introducing new challenges in effective system design and interaction as, for instance, effective prompting of LLMs can often replace the need to train a custom text-classification model. These advanced learning modalities have implications for the type of data and resources required to develop and deploy AI systems, and will be discussed in the next sections.

Fundamental Concepts

The development and deployment of AI systems are built upon several core concepts including encompass data, training, inference, and evaluation. Such concepts are key to understanding issues such as AI model evaluation, hardware considerations, and system design, which we discuss in the present report.

Data are a prerequisite for AI models to produce valid outputs, and the quality and quantity of data significantly impact AI performance. Data come in various forms: structured (like databases), unstructured (such as text and images), or semi-structured,

² Generalization refers to the model’s ability to apply the knowledge gained from a single example to recognize other instances of the same class. Specificity refers to the model’s ability to accurately identify the unique characteristics of the new class from that single example without confusing it with previously learned classes.

including JavaScript Object Notation³ (JSON) and Extensible Markup Language⁴ (XML) formats. While the possibility to rely on diverse and representative data is crucial for model performance generalization, data collection and use may pose important challenges related to privacy, bias, and other ethical considerations (Gebru et al., 2021).

The emergence of the Generative AI paradigm has transformed the data landscape. Many applications now utilize existing models pre-trained by large AI labs, reducing the need for organizations to train models from scratch. Simultaneously, synthetic data (see Squicciarini et al. (2024) for a discussion on synthetic content) have gained prominence, becoming more accessible through generative models. These advanced systems enable data generation either by extracting information from previously unstructured sources or by synthesizing entirely new data. Synthetic data can help augment limited datasets to improve model performance and contribute to preserving privacy by replacing sensitive information with realistic but artificial data. They can further help create balanced datasets to mitigate bias and improve model fairness, simulate rare events or edge cases – i.e. problems or situations occurring only at extreme (maximum or minimum) operating parameters –, thus enhancing model robustness. They may further facilitate development and testing in heavily regulated industries, where the use of real data is difficult or restricted.

Training involves that AI models learn from data and adjust their parameters to improve performance on given tasks. Various training methods exist, each suited to address different types of problems. Training processes face several challenges, including overfitting, whereby a model excels on training data but performs poorly on new, unseen data (Hoffer et al., 2017; Srivastava et al., 2014). Other significant hurdles include the need for substantial computational resources, access to relevant and high-quality data, and the requirement for specialized data training skills within an organization.

While training remains a core component of traditional machine learning, its role has evolved in the context of Generative AI and LLMs. For these advanced systems, training often takes the form of fine-tuning pre-trained models for specific applications, leveraging the power of transfer learning to adapt general knowledge to specialized tasks. This shift reduces the reliance on extensive in-house training, while nevertheless introducing new challenges in effectively adapting existing models to new use cases.

Inference involves applying a trained model to new, unseen data to make predictions or decisions. Inference can occur in real-time, as seen in recommendation systems, through batch processing for large-scale data analysis or whenever a chatbot system produces a reply. When deploying models for inference, key considerations include speed, accuracy, and resource efficiency and deployment options range from cloud-based solutions to edge computing or on-device processing, each having its own trade-offs and benefits (Irugalbandara et al., 2024).

It is important to note that models typically do not learn during inference: for example, when interacting with a chatbot, the underlying model typically does not learn from the conversation. Instead, each interaction triggers an inference process where the model processes the input and generates a response based on its training. In multi-turn interactions, such as ongoing conversations, new input is often concatenated with previous interactions to maintain context. Moreover, the inference landscape differs between traditional machine learning and the GenAI paradigm. Traditional models often require local inference or deployment by the organization that trained the specialized model. For instance, a real estate agency might deploy its own model to predict potential prices for sales listings. In contrast, GenAI models are frequently developed and hosted by commercial AI labs. Users of these services pay for inference based on the volume of input and output, often measured in tokens for text based tasks, or the number of generations for tasks like image, sound, or video content creation.

In 2024, the capabilities of “open weights” foundation models – i.e. models featuring widely available weights, which enable greater customization and deeper inspection at the expense of monitoring and moderation possibilities - have notably advanced, enabling companies to deploy robust applications on their own infrastructure. As we will explore further in this report, although this is feasible, the associated overhead — including expenses for hardware, maintenance, and the technical skills required — combined with decreasing costs of accessing models from major providers, often makes relying on in-house systems less appealing. This is unless there's a critical need for systems to operate locally.

³ JSON (JavaScript Object Notation) is a standard text-based format that represents structured data based on JavaScript object syntax and is commonly used to transmit data in web applications (e.g., sending data from the server to a client, for it to be displayed on a web page).

⁴ XML (Extensible Markup Language) is a markup language designed for encoding documents in a format that is both human-readable and machine-readable, often used for data storage and transport.

Tracing the Evolution of AI (Model) Capabilities

Advancements in hardware, AI model architectures, and training routines all influence model capabilities. However, while models are key enablers, it is ultimately application architects who determine which parts of performed tasks are routed to AI models versus traditional deterministic software systems, and how success and quality is defined. We here focus on AI model capabilities to understand the potential risks involved and how AI capabilities emerge and evolve. To this end, we transition from basic AI building blocks to examining the concrete evolution of AI's abilities, tracing the significant leaps that have redefined its boundaries. In particular, adding to the previous part on architectural information, we here concentrate on discussing the emergent functionalities and enhanced performance of successive AI model generations. Table 2 provides a detailed overview of AI approaches, highlighting their key capabilities, challenges, and ethical and societal implications, offering a foundation for understanding how AI model generations evolve in functionality and impact.

In what follow we explore the key trends that have characterized the field, and look at:

- (1) **Performance Enhancement**, intended as increasing performance in task completion.
- (2) **Generalization** across diverse domains.
- (3) **Multimodality** in data processing.
- (4) **Abstraction** and the move towards higher levels of abstraction in problem-solving.
- (5) **Flexibility and Modularity** intended as the growth of flexibility, adaptability in handling varied scenarios, and modularity enabling integration into more complex systems.

These trends reflect a fundamental shift from capabilities explicitly engineered into systems to capabilities learned and integrated by the models themselves, with AI that has progressed from simple classification and regression to complex reasoning and autonomous action. Shedding light on how AI has achieved its current capabilities is essential to better understand the products and services that have been put on the market recently and the exposure of different occupations to being automated – in part or in full – by AI systems, in addition to anticipating the potential directions for future innovation.

Traditional Machine Learning Capabilities

Early machine learning (ML) models primarily centered on classification and regression tasks, laying the groundwork for contemporary AI applications. Classification, i.e. the process of categorizing data into predefined classes (e.g., sorting emails into “spam” or “not spam”), and regression, i.e. the prediction of continuous values (e.g., predicting house prices based on various features), served as the cornerstones of decision-making and predictive analytics. These fundamental techniques remain ubiquitous across many sectors, and are employed for tasks such as identifying fraudulent transactions, diagnosing medical conditions, and forecasting financial trends. For example, relatively simple statistical models like logistic regression are commonly used to predict customer churn, while support vector machines⁵ have been utilized for image classification – often with varied success.

Feature engineering – i.e. the process of selecting and transforming raw data into formats that ML models can effectively use, in a way that is similar to translating raw information into a standardized format - remains a critical human element in traditional ML systems (Domingos 2012). While modern AI increasingly automates this process, understanding feature engineering helps explain both the capabilities and limitations of traditional ML systems: it is not merely about technical transformation; feature engineering also entails defining which variables should be excluded from the training set to avoid unintended bias. For instance, excluding gender or ethnicity in a financial product model aims to prevent biased outcomes. Moreover, selecting the most appropriate algorithm from a plethora of options often necessitated a trial-and-error approach, coupled with complex model validation procedures.

⁵ Support Vector Machines (SVMs) are a type of machine learning algorithm used for classification and prediction. They work by identifying an optimal boundary (hyperplane) to separate different categories of data, allowing for effective categorization of new, unseen data.

Table 2: Overview of AI Approaches

| AI APPROACH | ERA | KEY CAPABILITIES | KEY CHALLENGES | ETHICAL & SOCIETAL IMPLICATIONS |
|--------------------|----------------|---|---|---|
| Rule-Based Systems | 1950s-1980s | <ul style="list-style-type: none"> Explicit logic Predictable, predefined decisions | <ul style="list-style-type: none"> Inflexible in new contexts Requires domain experts for updates | <ul style="list-style-type: none"> May encode societal biases via poorly designed rules Limited adaptability to change |
| Machine Learning | 1990s- Present | <ul style="list-style-type: none"> Pattern recognition Task Specific predictions and classifications | <ul style="list-style-type: none"> Heavy reliance on high-quality labeled data Limited interpretability of model outputs | <ul style="list-style-type: none"> Risk of bias and discrimination from training data Lack of explainability, erodes trust |
| Deep Learning | 2012-Present | <ul style="list-style-type: none"> Complex pattern recognition Handles unstructured data (text, images, etc.) | <ul style="list-style-type: none"> High computational demands Vulnerable to adversarial attacks Difficult to interpret | <ul style="list-style-type: none"> Privacy concerns with large-scale data Potential misuse in surveillance or manipulation |
| Transfer Learning | 2015-Present | <ul style="list-style-type: none"> Leverages pre-trained models for new tasks Reduces need for large datasets | <ul style="list-style-type: none"> Potential bias transfer from source models Limited clarity on model transfer boundaries | <ul style="list-style-type: none"> Amplifies existing biases in pre-trained models Unintended impacts in new domains require careful validation |
| Generative AI | 2018-Present | <ul style="list-style-type: none"> Generates content (text, images, audio, code) Few-shot/zero-shot learning | <ul style="list-style-type: none"> Resource Intensive Difficult to control or constrain outputs Misuse (e.g., misinformation) | <ul style="list-style-type: none"> Misinformation risks (deepfakes) Potential displacement of human labor Content ownership dilemmas |
| Agentic AI | 2023-Present | <ul style="list-style-type: none"> Autonomous reasoning and planning Multi-step task execution | <ul style="list-style-type: none"> Accountability gaps in autonomous decisions Challenging to predict and control agent behavior | <ul style="list-style-type: none"> Reduced human oversight Risks of unintended consequences Threats to transparency and accountability |

Source: Authors' own compilation

Despite their continuing importance, advancements in AI are progressively diminishing the need for manual feature engineering, shifting focus towards models that can learn directly from raw data. This transition signifies a substantial conceptual development in model capability, shifting from human-crafted features to autonomous model learning.

Another fundamental approach, unsupervised learning, serves different but equally important functions in AI systems. Unlike supervised systems that learn from labeled examples, unsupervised learning discovers patterns and structures within data automatically. Imagine organizing a library of books without predetermined categories — the system finds natural groupings based on the books' characteristics. A key application of this is data compression through dimensionality reduction, which can be thought of as distilling complex information into its essential elements. This process enhances the visibility of underlying patterns and has also been implemented as part of recommendation systems, i.e., identifying consumer preferences based on prior purchase patterns.

Another significant area is clustering, which entails grouping similar items together, aimed to identify groups of observations with shared characteristics, often accompanied by an explicit representation of those characteristics (Aggarwal and Reddy, 2014). In fact, many sections of the forthcoming "UNESCO AI Landscape Report" leverage a combination of both dimensionality reduction and clustering techniques. Unsupervised learning acts as a critical mechanism enabling us (the authors) to identify and hierarchically group AI discourses. This approach facilitates the synthesis of core narratives from hundreds of sources, enabling a more comprehensive and structured mapping of complex information.

Deep Learning Breakthroughs

Deep learning has significantly advanced AI capabilities, altering the scope of what machine learning systems can achieve, particularly in terms of growing complexity for both inputs and outputs. The introduction of architectures such as Convolutional Neural Networks (CNNs, i.e. specialized for processing visual patterns, similar to how human vision works) (LeCun et al., 2015), Recurrent Neural Networks (RNNs, which are designed to process sequences of information, like understanding sentences) (Hochreiter and Schmidhuber, 1997), and Transformers (i.e. advanced systems that can process and generate human-like text) (Vaswani et al., 2017) has been pivotal in this development, enabling breakthroughs across numerous domains. CNNs allowed AI to perceive and interpret visual information with unprecedented precision. RNNs, conversely, have unlocked new possibilities in sequence modeling, making it feasible to process and understand sequential data such as time series and natural language. The advent of Transformers revolutionized natural language processing, with models like BERT (Devlin et al., 2018) and GPT achieving markedly superior language understanding and generation. These models are now capable of not only processing text but also generating coherent and contextually relevant content.

Furthermore, deep learning has drastically improved both model accuracy and scalability, enabling AI to handle considerably larger and more complex datasets, thus moving from processing thousands of examples to dealing with billions of information pieces. This enhancement has unlocked the potential for training models on web-scale data, achieving human-level performance on a variety of language-based tasks that were previously unfeasible. This dramatic increase in capability raises new policy challenges around data usage, compute resources, and model access.

The capacity to process vast amounts of data, coupled with improved scaling, has expanded the boundaries of what AI can accomplish. This increased scale is exemplified by large language models trained on massive datasets to deliver human-like capabilities in language-based tasks (Brown et al. 2020). Through deep learning, AI has not only become more accurate but also more versatile and capable to address the intricate challenges posed by complex data and diverse applications. We have transitioned from models capable of processing predefined tabular feature groups to models able to process images, signal, and text sequences, and to generate more complex outputs — not only classes and number estimates — thereby enabling new types of applications across different domains.

Generative Abilities and Multimodal Integration

The emergence of generative AI models has marked a significant turning point, expanding the capabilities of AI systems to create high-quality text, images, and other forms of synthetic content (Squicciarini et al. 2024). This development signals a fundamental shift in how humans interact with AI, moving beyond analysis to creation.

Initially, between 2018 and 2020, generative AI focused primarily on text generation, with models such as OpenAI's GPT series (starting with GPT-1 in 2018 and progressing to the present day) demonstrating this new potential by generating diverse outputs including articles, poetry, and even computer code (Radford et al., 2019). These advances blurred the lines between human and machine-generated work. Then, from 2021 to 2022, AI capabilities expanded significantly into image generation, with models producing increasingly photorealistic images from textual prompts (Ramesh et al., 2021). This ability, alongside the ability to create increasingly sophisticated video content, began challenging established domains, including art, advertising, and entertainment. These capabilities captured widespread public attention, mainstreaming AI through accessible tools like ChatGPT, which brought generative AI into everyday use.

Feature Focus 1: Multimodality

Input:



Audio recording provided in German saying: "How many flowers can be seen – answer in English"

► Model: There are four flowers on the picture.

Images Source: © Quang Ho/ Shutterstock.com*

The year 2024 marked the arrival of multimodal AI. Google's Gemini 2.0 Flash Experimental (introduced in December 2024) exemplifies this, with its ability to simultaneously process image and audio inputs. These models are now trained to handle various types of input, including live video and audio streams, enabling interactive dialogue with the model. This capability opens a vast array of applications but also introduces new challenges. What are the implications of deploying such systems in public spaces and integrating them into everyday gadgets? Theoretically, such models could perform surveillance – streaming a video feed from a webcam and automatically triggering actions (e.g., sending notifications or alerting authorities) upon detecting specific situations. These possibilities underline the need to carefully consider the ethical implications of such developments, especially regarding privacy and the potential for misuse.

Feature Focus 1: Multimodality cont.

Input:



You are an intelligent surveillance assistant analyzing images from an apartment in real time. Each input is a single image, and your task is to detect and classify noteworthy events or objects in the image. Provide the output as a JSON object with the following structure:

JSON Structure:

```
{
  "people": [ // List of descriptions of people detected.
    "e.g., 1 adult male",
    "2 children"
  ],
  "pets": [ // List of descriptions of pets detected.
    "e.g., 1 dog, black Labrador"
  ],
  "suspicious_activity": [ // List of suspicious activities or unusual behaviors.
    "e.g., person near locked drawer",
    "window opened"
  ],
  "unusual_objects": [ // List of unexpected or out-of-place objects.
    "e.g., box near door",
    "large suitcase"
  ],
  "lights_status": Boolean, // Whether the lights are on (true)
  "doors_status": [ // Status of doors in the image.
    "e.g., front door open"
  ]
}
```

► OUTPUT

```
{
  "motion_detected": true,
  "people": [
    "1 adult person"
  ],
  "pets": [],
  "suspicious_activity": [],
  "unusual_objects": [],
  "lights_status": true,
  "doors_status": [
    "front door open"
  ]
}
```

Images Source: © GBJSTOCK/ Shutterstock.com*

Key Takeaways:

- AI can now process and integrate multiple types of input (audio, image, text) simultaneously
- Cross-language and cross-modal capabilities enable new applications
- Real-time monitoring capabilities raise significant privacy and ethical concerns
- Structured outputs (like JSON) enable integration with other systems
- New governance frameworks needed for example for multimodal AI deployment in public spaces

Building upon these initial generative capabilities, and particularly between 2023 and 2024, AI systems have advanced towards integrating and processing multiple types of data inputs and outputs – the so-called multimodality, i.e. the ability to process and combine different types of information such as text, images, audio, - in ways that mirror human cognitive integration. Previously, many AI models were limited to single-modality outputs. This meant that they were trained on, and could only produce, a single type of data (e.g., only text, or only images). Now, models can process and combine data streams such as audio, video, images, and text. For example, models can now process spoken language to hold near-instantaneous conversations, or simultaneously translate between languages. This allows AI applications to move past single-input applications to those that combine diverse data, enhancing contextual understanding. This integration of multimodality broadens AI's scope in the same way it poses new challenges.

Code Generation and Automation

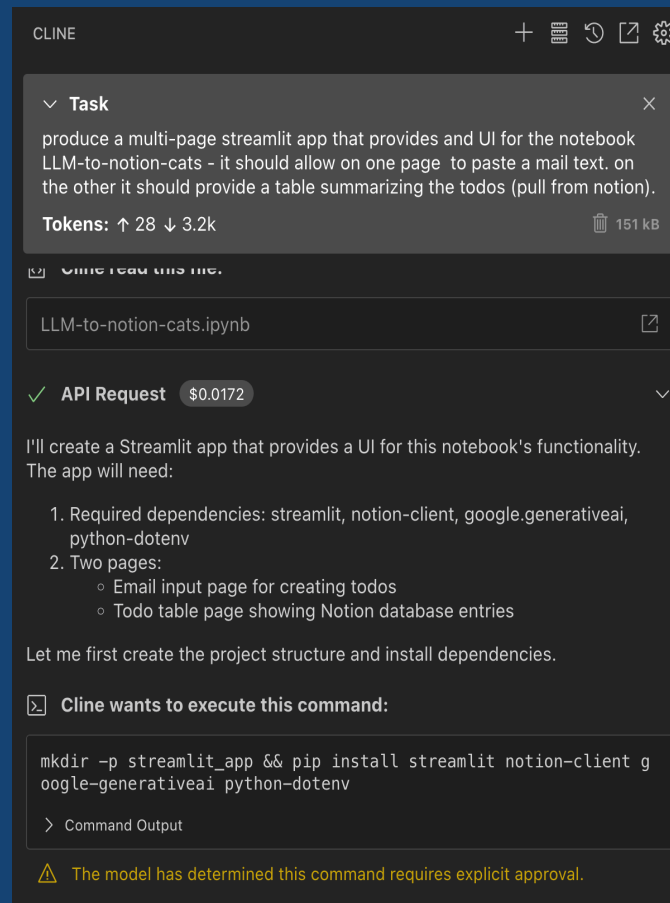
A significant advancement in AI capabilities is the ability to generate outputs that can be objectively verified. Unlike natural language generation, where quality assessment remains largely subjective, these new capabilities produce testable, deterministic outputs. Code generation represents a prime example of this capability, as the output can be executed, tested, and verified against objective criteria (Chen et al. 2021). Such a shift marks a crucial shift from AI systems that generate plausible-seeming content to those that can produce demonstrably correct solutions.

AI models are increasingly capable of generating and manipulating code, impacting software development practices. This progression includes the ability to produce structured, machine-readable outputs suitable for use as inputs by other systems (e.g., an LLM generating a structured list of ingredients that a program could process). This capability extends to the generation of explicit and deterministic representations, such as code that performs a specific calculation, allowing verification through execution and testing. Initially, AI models focused on completing code snippets, but now they can generate entire programs from natural language descriptions. This has the potential to shift software development focus towards higher-level design, system architecture and specification, while relegating lower-level implementation details to AI. These models also carry the ability to debug code by identifying and correcting errors, thereby streamlining the development and verification process. AI-driven tools are also contributing to software testing, refactoring, and maintenance, by automatically generating test cases, identifying bugs, and improving code.

This shift towards verifiable outputs has significant implications for AI governance and deployment, as it may enable the definition of objective quality metrics and automated verification systems; the establishment of clear performance boundaries; and their reliable integration into critical systems.

The increasing capacity of models to produce executable code that can be tested and verified through defined metrics of quality is significant for raising overall reliability and trust in AI technologies in the first place. Current research emphasizes the use of stochastic AI models to generate deterministic code, reflecting a need to enhance the robustness and predictability of AI-generated software (Jimenez et al., 2024). This evolution highlights the expanding role of AI not only in generating code, but also automating aspects of its verification, in turn raising questions about who controls and ensures the integrity of AI-generated code at scale. While these advancements could potentially democratize software development, they also raise questions about skill obsolescence and what constitutes expertise in this changing landscape, along with the challenges of maintaining AI-generated software over time.

Feature Focus 2: Coding Agent



Images Source: Screenshot from experiment by Authors

The latter half of 2024 has witnessed the rise of sophisticated open-source coding agents capable of autonomously planning and executing software development tasks. A prime example is "Cline," with its ability to create a multi-page application based on the "LLM-to-notion-cats.ipynb" notebook. The agent systematically breaks down the requirements, detailing the project structure and necessary dependencies while specifying a two-page architecture, despite receiving imprecise instructions with grammatical errors: one page dedicated to email text input and another for to-do table display. Upon understanding the requirements, Cline formulates an initialization command to create the project directory and install essential Python packages. This emerging technology represents a significant advancement in automated software development, where agents can effectively translate natural language descriptions into concrete development steps, managing everything from code generation to dependency handling and architectural planning. While these agents look promising in streamlining the development process, particularly for straightforward projects, they maintain a safety-first approach by requiring explicit user approval for executing commands. This development suggests an evolving landscape in software development methodologies, where AI agents increasingly support the technical and planning aspects of project implementation.

Key Takeaways:

- AI can now autonomously plan and execute software development tasks
- Systems can interpret imprecise human instructions into structured technical plans
- Safety measures (like requiring user approval) are being built into autonomous systems
- The role of human developers is shifting toward high-level design and oversight
- Raises questions about the future of software development practices, jobs and skills

Reasoning, AI Agents and Autonomy

The capabilities discussed so far enable increasingly flexible automation that leverages engineered systems built around Large Language Models (LLMs). A growing trend is represented by the integration of these capabilities directly into the models themselves, also reflected by the emergence of agentic systems, which represent a shift toward AI capable of executing complex

Feature Focus 3: Agentic Computer Use

Note: This example has been edited and simplified for presentation in this report format. While the original system operates with actual computer interfaces, mouse movements, and screenshots, we can only show the textual representation here. The example preserves the essential structure and nature of how these models verbalize their “thinking” and actions during autonomous computer operation.

USER: Please add the following 5 customers from the provided spreadsheet into the CRM: Company 1, Company 2, Company 3, Company 4, and Company 5. List & verify each customer's information as you input it into the CRM.

▶ **Agent:**
Get oriented

I'll help you add these 5 customers into the CRM. Let me start by taking a screenshot to confirm the spreadsheet and CRM are both visible.

🖼️ Screenshot

Switch to CRM tab I see the spreadsheet is active.
I need to switch to the CRM application to start adding customers. First, I'll click on the CRM tab.

🖱️ Move to 600, 50
🖱️ Left click
🖼️ Screenshot

Begin Customer Input

Great, I've switched to the CRM.

Now, I will start adding the customer information one by one, starting with 'Company 1'

Recent examples of generalist autonomous agentic systems, such as Claude Computer Use by Anthropic, have emerged in late 2024, pointing to a departure from narrow, task-specific agents. These systems are designed to interpret broad instructions and, like a human employee, utilize the computer interface – keyboard, mouse, and screenshots for contextual awareness – to execute tasks. This approach has immense potential for automating processes, especially within legacy systems where direct integrations are challenging or impossible. By mimicking user interaction, these agents can navigate and operate software as a human would.

This new paradigm, while promising, raises critical considerations:

- **Security and Safety:** Ensuring the secure operation of these systems is paramount. Questions remain around access control, data protection, and the prevention of unintended consequences.
- **Identification and Access Control:** How should these agents be identified and authenticated? What level of access should they be granted to different functions and data within the system?
- **Data Privacy:** Defining the parameters of acceptable data access and minimizing data exposure are crucial. What data should these systems be allowed to view and interact with?
- **Unintended Applications and Purpose:** Somewhat concerning is the potential for these systems to circumvent intended purposes. For example, online discussions have rapidly identified an illustrative, if somewhat cynical, “use case” whereby these systems could be used to automate the viewing of Corporate Online Compliance courses, automatically progressing through videos and completing multiple-choice questionnaires. This serves to highlight how automation can actually bypass the goals of the systems in place today.

Key Takeaways:

- AI agents can now operate computer interfaces like human users, enabling automation of legacy systems
- This capability brings both opportunities (process automation) and risks (security, access control)
- Current systems can bypass intended human-centric processes, highlighting the need for new governance approaches

tasks based on high-level goals. This evolution can be seen as a progression from basic reasoning capabilities to integration with action (agency) and eventually to multi-agent coordination. Each level brings new capabilities and challenges.

In terms of base reasoning capabilities, this involves integrating a form of reflective inference (often referred to as Chain-of-Thought), enabling the model to explore and evaluate multiple potential outputs against a task definition by reasoning through different possible scenarios and outcomes (Wei et al., 2023). This reflective inference approach is combined with the paradigm of

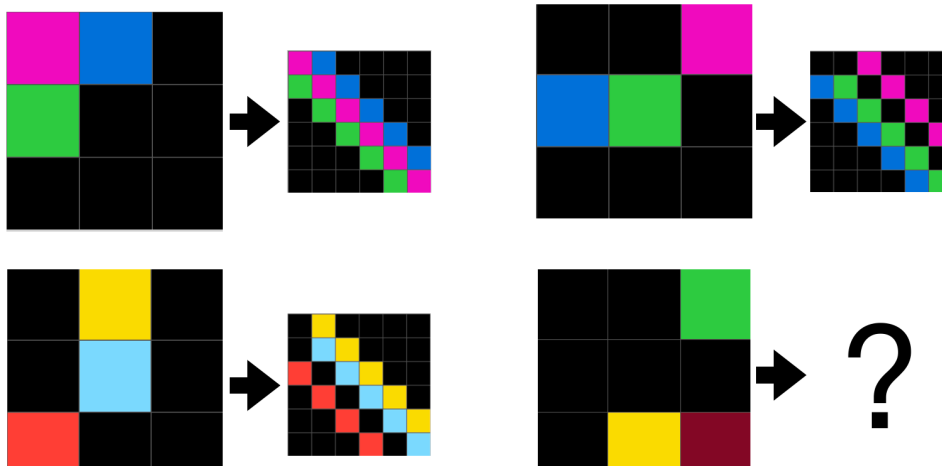


Image source: Compiled by the author using benchmark data presented in Chollet, F. (2019). On the measure of intelligence. arXiv. <https://arxiv.org/abs/1911.01547> provided under the Apache 2.0 license <https://github.com/fchollet/ARC-AGI>

The ARC-AGI benchmark, featuring abstract visual puzzles like this one from Chollet (2019), tests AI models' general intelligence capabilities. In December 2024, OpenAI's o3 model achieved breakthrough performance on this benchmark, surpassing 75% on the semi-private evaluation set - a score comparable to average human performance, though still below the near-perfect scores achieved by STEM graduates.

"inference-time compute" (also known as "test-time compute") (OpenAI, 2024). This paradigm involves generating and probing different possible outcomes, often in a hierarchical manner following a pre-set plan, to select for the optimal outcome of a certain task. This can be thought of as the system trying out different solutions and choosing the best one. This capability is particularly important for tasks with objective, verifiable answers, enabling their integration into automated systems rather than relying solely on human qualitative evaluation. The combination of reflective inference and the test-time compute approach are crucial elements of more advanced agentic systems.

Agentic systems, which integrate reasoning with the ability to take action, move beyond simple prompt responses, incorporating elements of planning and tool use to execute multi-step processes with reduced human intervention. For example, a simple agentic workflow might involve one agent monitoring an email mailbox to identify to-do items, create calendar entries or other such tasks. Another agent may then review the to-do list based on a predefined schedule, providing daily summaries or reminders to the user. A third agent may then monitor outgoing emails and cross off tasks if they have been "handled off" - highlighting a more complex example of inter-agent interactions.

Late 2024 saw examples of these capabilities, including multimodality, demonstrated in preview systems that control computers through screen analysis, mouse control, and text input. This ability to directly control a computer system via AI, raises significant possibilities – such as the automation of legacy systems – while also triggering potential challenges stemming from the risks related to errors, reduced human oversight, malicious use, and the current technical limitations in handling diverse or noisy inputs.

Multi-agent coordination, where AI agents communicate and coordinate to achieve shared objectives, is an ongoing area of research. Current implementations are primarily in simulation or controlled settings. Real-world applications face challenges related to unpredictable scenarios and the need for robust error handling. It is important to note that while these systems can perform increasingly complex tasks, they still often rely on clearly defined objectives and face limitations in handling unforeseen events and diverse, noisy inputs.⁶

⁶ "Noisy inputs" refer to data that is inaccurate, incomplete, ambiguous, or irrelevant, hindering the ability of AI models to learn effectively and make reliable decisions. This is a data quality issue, distinct from the diversity of viewpoints and backgrounds.

Model Evaluation

Assessing how well an AI model or AI-powered application performs and whether they can adapt to new situations is crucial before using them in the real world. This process, known as model evaluation, involves testing the AI in scenarios the system has not seen before. This is similar to what a student can be asked to do in an exam: only an exam reveals how well a student has learned the material and masters the task. Similarly, model evaluation reveals how well an AI can generalize its knowledge along measures such as accuracy, precision, and recall. Techniques like cross-validation (splitting the data into multiple subsets to train

Feature Focus 4: Reasoning Models

Amina has four daughters, and each of her daughters has a brother. How many children does Amina have?

Thought for 11 seconds

► Alright, let's tackle this problem step by step. So, the question is: Amina has four daughters, and each of her daughters has a brother. How many children does Amina have? Hmm, okay. Let's parse this out. First, Amina has four daughters. That part seems straightforward. So, that's four girls. Then, each of these daughters has a brother. Now, here's where I need to be careful. The key point is that each daughter has a brother. But does that mean each daughter has their own brother, or do they share a brother? If each daughter had their own brother, that would mean each of the four girls has a different brother. So, that would be four brothers, one for each daughter. But wait, if that's the case, then the total number of children would be four daughters plus four brothers, totaling eight children. But that seems like a lot, and maybe I'm misunderstanding the question. [...]

Amina has four daughters, and each daughter has at least one brother. Since the brother is shared among all the daughters (they are siblings), there is only 1 son.

Total children = 4 daughters + 1 son = 5 children.
The answer is 5.

The autumn of 2024 witnessed the emergence of sophisticated reasoning models, with OpenAI's o1 and subsequent offerings from other companies leading the way. Notably, the Qwen group under Alibaba Cloud released QwQ, an openly available model and on January 20th DeepSeek released their R1 – riveling the performance of OpenAI's o1. These models employ training techniques that explicitly encourage the generation of “reasoning traces” – essentially, the step-by-step logic – before delivering a final answer. This approach offers dual benefits. First, it empowers these models to consider diverse possibilities and alternatives (through “chains of thought”), enhancing the accuracy of responses, particularly in complex scenarios requiring deeper analysis. Second, the accessibility of these reasoning traces may allow for a better understanding of how these language models reach their conclusions, which is critical for establishing trust and identifying potential flaws. However, this transparency is not always available; several frontier model providers have chosen not to expose these reasoning traces, citing concerns about competition and the potential for competitors to use this output for training purposes. Despite this limitation, this improved reasoning performance comes with a significant trade-off: it requires substantially more compute during the inference process. This increased computational cost needs careful consideration when deploying and scaling these models.

Key Takeaways:

- Modern AI can show its reasoning process, making decisions more transparent and verifiable
- Models can explore multiple possibilities before reaching conclusions
- There's a tradeoff between detailed reasoning and computational costs
- Some providers keep reasoning processes private for competitive advantage, raising transparency concerns

and evaluate the model on different parts) and holdout sets (reserving a portion of the data exclusively for final model evaluation after training) help ensure the model is ready for real-world deployment.

However, evaluating AI models may at times be cumbersome, especially in relation to different real-world scenarios and treating all (sub)groups fairly, and GenAI applications have made model evaluation even more complex (Ribeiro et al., 2016; Chang et al., 2024). Advanced models require assessment across various areas, including how well they understand facts, their ability to reason, and whether their decisions are fair. As evaluating these models can be difficult, alternative methods are sometimes used. These include the “Elo” scoring system, known for ranking chess players, which can be used to compare the performance of models. Also, human evaluators often compare model outputs directly. This is particularly useful in relation to tasks involving human language and creativity, where it can be difficult to rely on traditional metrics. In the upcoming “UNESCO AI Landscape Report”, we will propose a more detailed discussion of evaluation approaches, evolution and various shortcomings, including evaluation methods, benchmarks, and the direction this field is heading towards.

The Hardware Layer

The hardware infrastructure supporting AI has evolved significantly alongside the models themselves. Traditional AI models are relatively resource-efficient, benefiting from years of development in both local and cloud infrastructure. Organizations have established practices for running these models effectively, largely addressing deployment challenges, although the complexities of Machine Learning Operations (MLOps) - i.e. the practices, processes, and tools used to deploy, manage, and monitor ML models in production environments - remain for those managing their own infrastructure. The hardware required for AI depends heavily on the chosen AI technology and learning modality. For example, generative AI models place far more significant demands on hardware compared to rule-based systems.

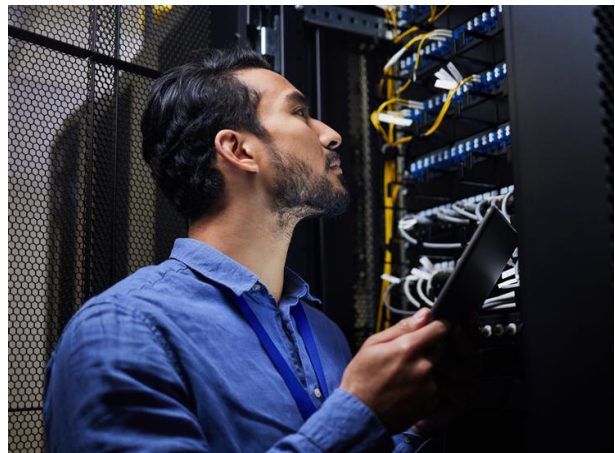


Image Source: © PeopleImages.com - Yuri A/Shutterstock.com*

A crucial distinction in AI hardware requirements lies between training and inference phases. Pre-training large language models demands immense computational resources, often requiring specialized machines with costs reaching hundreds of millions of US dollars for the most advanced models (Cottier et al., 2024). This scale of operation is typically the domain of major technology companies and research institutions. However, the landscape becomes more diverse when considering fine-tuning, a process of adapting pre-trained models to specific tasks. Instruction finetuning, which is the process of adapting pre-trained models to perform specific tasks by providing instruction data that guides the model on the desired outputs for given inputs, can be undertaken by mid-sized companies with substantial but not prohibitive resources. At the other end of the spectrum, smaller models can be fine-tuned by hobbyists on more modest hardware, showcasing a more accessible edge to AI development.

The evolution of AI hardware has seen a shift from traditional Central Processing Units (CPUs) to GPUs (Graphics Processing Units) and specialized AI accelerators like Tensor Processing Units (TPUs). These hardware innovations have dramatically improved the speed and efficiency of both training and inference processes. Recently, several companies have begun developing custom chips optimized specifically for LLM inference, aiming to reduce the delay before a transfer of data begins following an instruction for its transfer, i.e. the so-called latency, and increase energy efficiency. This trend towards specialized hardware reflects the unique computational patterns of LLM architectures and the growing demand for efficient, high-performance AI inference capabilities.

The inference stage, where trained models are applied to new data, can rely on a range of hardware options. Remote inference via LLM providers has become a popular choice, offering scalability without requiring in-house hardware management. This model of inference is particularly common for generative AI paradigms, due to the computational resources required for both the training and inference of these complex models. At the same time, self-hosted inference is becoming more accessible, allowing organizations to maintain control over their AI operations.

A notable trend throughout late 2023 and into 2024 is the rise of small, efficient models capable of running on local machines or edge devices (devices like smartphones, sensors, or embedded systems that process data closer to where it is collected), bringing AI capabilities closer to end-users. Various techniques have emerged to improve hardware efficiency, often at the cost of some performance though. These optimizations are particularly relevant when the full capabilities of a larger generative LLM are not necessary for a specific use-case.

This trend towards efficiency reflects a maturing understanding of the tradeoffs between model size, performance, and resource requirements in the AI hardware landscape. In the latter part of 2024, an opposite trend became apparent with the introduction of “reasoning models” and the expansion of compute during inference time. This evolution in generative models extends inference compute time to mimic cognitive processes. These models, in practice, produce extensive text traces exploring various directions and contemplating diverse outcomes through an internal monologue. This approach has demonstrated enhanced performance in complex tasks that require thoughtful deliberation. From a hardware and resource perspective, this necessitates increased resources at inference time, rendering specialized hardware more pertinent, but also escalating operational costs and complicating their deployment on less sophisticated local infrastructure. Energy consumption also remains a significant concern, particularly for large-scale AI operations, and is driving ongoing research into more energy-efficient hardware and model architectures.

AI Systems, Design, and User Interfaces

Users typically experience artificial intelligence through engineered systems, not directly through AI models. This makes it important to explore the role of system design in translating AI capabilities into practical applications, and emphasizes the relevance of user-facing designs (Sculley et al., 2015). AI models are integrated into broader systems that govern their functionality, set their limitations, and condition how they are presented to end users. These systems encompass crucial components such as data storage protocols, output supervision mechanisms, and input filtering processes. The design choices embedded within these systems exert a significant influence on user experience and ethical considerations.

Between the AI model and the user lies a layer of engineered systems, where designers make explicit decisions about user interaction and system behavior, including on data storage policies, output supervision, and input filtering. System designers must consider whether and how to store user conversations, implement secondary systems to ensure generated output aligns with specified guidelines, and integrate specialized models to filter potentially malicious inputs. As an example, consider a healthcare system designed to assist physicians in their daily routines. In this system, the act of a physician logging into their daily schedule could automatically trigger a complex series of AI processes. First, patient data from the Electronic Health Record (EHR) would be aggregated and summarized, drawing on multiple models to ensure the factual accuracy of the summary. A separate model would then be employed to refine the output, making sure that it is concise and directly relevant to the physician's immediate needs for their upcoming consultation. This would aim to ensure that the summary only includes necessary information such as past diagnoses, medications, and relevant patient history. The user interface might further include buttons allowing the physician to quickly access additional details regarding specific conditions or past medications, allowing her to interact with the AI model in a controlled manner and to gain deeper insights on demand. In this case, the user does not directly interact with the AI but benefits from its outputs and actions, highlighting the importance of a well-engineered system.

In addition, as AI systems become more complex, they often incorporate multiple models that need working synchronically, requiring careful orchestration to achieve the desired outcomes. This multi-model approach allows for more sophisticated functionality while at the same time increasing the complexity of system design and management. Consequently, many ethical considerations in AI use extend beyond the development of individual models and encompass the entire AI system. The responsibility for ethical AI implementation increasingly falls on the developers of AI-enabled products or systems, rather than solely on the creators of individual AI models (Mittelstadt et al., 2016).

The landscape of AI system design has undergone a significant transformation. While traditional machine learning techniques often resulted in systems with limited complexity, the flexibility of generative systems and emerging multi-modality has opened doors to increasingly sophisticated architectures. AI models are becoming integral components of complex systems, enabling more advanced and nuanced applications with sophisticated interfaces. A prime example is the creation of a multi-speaker podcast from an uploaded PDF document. This process involves multiple AI models working simultaneously: one model may identify key points in the PDF and draft a manuscript, which is then passed to another model for speech synthesis and audio production. The end user interacts only with the final product, unaware of the intricate AI processes behind the scenes. While chat interfaces have mainstreamed AI applications, the scope of integration is expanding, and interfaces are diversifying.

Even more advanced and sophisticated applications are likely to emerge as AI models improve in stability, produce fewer errors that could hinder integration, and better approaches to error checking and mitigation are developed. The evolving landscape of AI systems and interfaces presents both opportunities and challenges for policymakers, necessitating a nuanced understanding of the interplay between technological capabilities, user experience, and ethical considerations related to AI development, deployment and governance.

The Co-Evolution of AI Capabilities and Governance Challenges: Some First Conclusions

As we trace AI's evolution from basic classification tasks to today's sophisticated systems, a crucial pattern emerges: understanding how AI capabilities develop technically is essential for grasping how they can (re)shape applications and governance's needs. Each advance in AI does not merely enable new uses: it fundamentally changes how we need to think about key issues such as privacy, oversight or accountability.

This evolution begins with the core technical building blocks, i.e. AI models themselves. As these models grow in sophistication, they influence how developers can apply their capabilities in specific contexts. The decisions that developers make for instance in relation to the specific tasks to delegate to AI, how to integrate these systems within existing structures, and how to measure success, directly shape the applications we see emerging. Yet the human component of AI systems gets often overlooked. It is the engineers designing these systems, the developers integrating them into broader applications, and the decision-makers wrestling with new challenges, among others, that decide what today and tomorrow's AI technologies look like and whether or not they are put at the service of humanity. This in turn determines the opportunities that arise as well as the challenges that different stakeholders including individuals, civil society, economic agents and policy makers will have to face and address, to avoid that AI overlooks or even worse violates human rights, human dignity and fundamental freedoms.



Image Source: © PeopleImages.com - Yuri A/Shutterstock.com*

Emerging applications invariably strain our existing frameworks for oversight. For instance, when AI generates images or code, it raises novel questions about verification and responsibility. When AI agents make decisions in complex environments, we need new ways to think about accountability. This is not merely about technical capabilities, but rather about understanding how these capabilities interact with human judgment and interact with or shape existing institutional structures and impact relevant agents.

This challenge is deepened by an interesting paradox in AI's technical evolution: as systems become more sophisticated, they increasingly internalize functions that were once explicit and visible and, to a great extent; traditionally, the realm of human decisions and tasks. This makes it harder to assess their behavior and verify their decisions. Some industries respond by demanding more explainability before adoption, while others prioritize performance gains despite uncertainty about risks. For example, an LLM application processing paperwork in the public sector could drastically improve efficiency while probably saving development costs yet explaining precisely why a document was routed in a specific way may prove exceedingly difficult or impractical. Conversely, an AI-enabled trading algorithm might achieve superior performance but operate with such complex logic that its decisions remain opaque – which is acceptable as long as it produces the expected (positive) outcomes. In both cases there are trade-offs, as enhanced performance often comes with reduced transparency and auditability. Neither approach is inherently right – each context requires its own careful balance. Such (new) issues call for a systematic exploration of the intricate relationship between AI technology, the ecosystem, emerging applications and the coevolving trade-offs that may emerge.

This is exactly the call that UNESCO Recommendation on the Ethics of Artificial Intelligence makes, when reminding about the need to abide by principles including proportionality and do no harm coupled with responsibility and accountability, safety and security, fairness and non-discrimination, right to privacy and data protection, and human oversight and determination, among others. This in turn calls for multi-stakeholder and adaptive governance and collaboration, that take into account shifts in technologies, the emergence of new groups of stakeholders, and allow for meaningful participation by marginalized groups.

In conclusion, this primer aims to foster a more informed and human-centred conversation on AI. It enables a better understanding of what AI technologies are and can do, what their building blocks are, and how they interact to create diverse functions and products that impact society, businesses, and every day's life. The ability to identify and address the specific elements that AI systems are made of and the interfaces between them defines how tasks are solved and helps identify those suboptimal or problematic system outcomes. Also, understanding the way AI engineers approach and understand technical building blocks, the type of capabilities that emerge and the way these all interact can facilitate a more informed and solution oriented dialogue among relevant stakeholders, to truly put AI at the service of societies.

Glossary

- **AI (Artificial Intelligence):** The broad field of computer science focused on creating systems that can perform tasks that typically require human intelligence, such as learning, problem-solving, and decision-making. AI systems aim to replicate human-like cognitive functions through computer algorithms and data analysis.
- **AI Agent:** An AI system capable of perceiving its environment, making decisions, and acting autonomously to achieve specific goals. These can range from simple automated tasks, like responding to emails, to complex planning and problem solving, like controlling a robot in a factory.
- **AI Accelerator:** Specialized hardware (like GPUs and TPUs) designed to drastically speed up and increase the efficiency of AI model training and inference by optimizing specific operations, like matrix multiplications. These are crucial for handling the massive computational demands of deep learning models.
- **Algorithm:** A step-by-step set of rules or instructions that a computer follows to solve a problem or complete a specific task. Algorithms are the underlying recipes that drive AI systems.
- **Attention Mechanism:** A technique used in neural networks, especially Transformers, that allows the model to focus on the most relevant parts of the input data when processing it. It allows the model to weigh the importance of different words, images, or other data in relation to each other.
- **Batch Processing:** A method of processing data in large groups or batches rather than individually in real-time. It's commonly used for large-scale data analysis where immediate results are not necessary.
- **BERT (Bidirectional Encoder Representations from Transformers):** A powerful, pre-trained language model developed by Google that revolutionized natural language processing (NLP). It understands context bidirectionally (from left to right and right to left).
- **Chain-of-Thought (CoT) Prompting:** A technique that encourages a language model to explicitly reason step-by-step, by generating a chain of intermediate thoughts or inferences before arriving at an answer. This can significantly improve the model's ability to solve complex or multi-step tasks by emulating a form of human-like reasoning.
- **Classification:** A supervised machine learning task that categorizes data into predefined classes or groups. Common examples include classifying emails as "spam" or "not spam", and categorizing images into "cat," "dog," or "bird."
- **Clustering:** An unsupervised machine learning technique that groups similar data points together based on their inherent characteristics, without prior knowledge of those groups. This technique is used for data exploration, pattern discovery, and customer segmentation.
- **CNN (Convolutional Neural Network):** A type of deep neural network particularly suited for processing visual patterns and image data. CNNs are often used in image recognition, image classification, and object detection.
- **CPU (Central Processing Unit):** The primary processor in a computer that performs the bulk of the machine's computing operations. While CPUs can be used in AI, they're less efficient than GPUs or TPUs for the parallel processing needed in deep learning.
- **Cross-Validation:** A model evaluation technique that involves splitting the data into multiple subsets, using some subsets to train the model and others to validate its performance. This helps ensure that the model generalizes well to unseen data.
- **Data:** Information used to train and evaluate AI models. This includes structured data (like databases), unstructured data (like text and images), and semi-structured data (like JSON files).
- **Data Augmentation:** A technique to artificially increase the diversity and size of a training dataset by creating modified versions of the original data. This involves operations like rotating, cropping, or adding noise to images, or paraphrasing text data, to improve the robustness of a model.



- **Deep Learning:** A subfield of machine learning that uses artificial neural networks with multiple layers (deep neural networks) to learn complex patterns from data, allowing AI to achieve higher levels of abstraction than traditional ML.
- **Dimensionality Reduction:** Techniques used to reduce the number of input features in a dataset while preserving essential information. This simplifies data processing, visualization, and reduces training complexity by removing redundant information.
- **Edge Computing:** Processing data closer to its source (e.g., on a user's device, or a local server) rather than in a centralized data center, reducing latency, bandwidth usage, and reliance on network connectivity. Edge computing is increasingly important for real-time applications like autonomous vehicles and IoT devices.
- **Elo Score:** A rating system commonly used in games like chess to assess and compare relative skill. In AI evaluation, it can be used to compare the performance of different models by playing them against each other in a standardized testing environment.
- **Feature Engineering:** The process of selecting, transforming, and extracting the most relevant features (inputs) from raw data to improve the performance of a machine learning model. Feature engineering can involve cleaning data, scaling it, or creating new features from existing ones.
- **Few-Shot Learning:** A machine learning approach that allows models to generalize effectively from a limited number of labeled examples, often 2-5 instances per class, which is highly beneficial when data is scarce.
- **Fine-Tuning:** The process of further training a pre-trained AI model on a specific dataset to adapt its knowledge to a new, related task. Fine-tuning leverages the existing capabilities of the pre-trained model, significantly reducing the required training data and computation compared to training from scratch.
- **Foundation Model:** A large AI model (often a large language model) trained on a vast amount of unlabeled data. These models serve as versatile foundations that can be adapted (fine-tuned) for a wide range of tasks, acting as a starting point for downstream applications.
- **Generative AI:** A class of AI models that can create new content, such as text, images, audio, or video, that was not explicitly present in its training data. This represents a shift from models that analyze data to those that can create new forms of data, which has implications for fields like art, design, and content creation.
- **GPU (Graphics Processing Unit):** A type of processor initially designed for graphics rendering, now widely used for AI due to its parallel processing capabilities. GPUs excel at matrix multiplication, which is a key computation in deep learning, enabling faster model training and inference.
- **Holdout Set:** A portion of the data that is deliberately kept separate from training and used solely for evaluating the final performance of the model, providing a realistic assessment of its generalization capability.
- **Inference:** The process of using a trained AI model to make predictions or decisions based on new, unseen data. This is the stage where the model is used in real-world scenarios to perform its intended task (e.g., generating a response in a chatbot).
- **Inference-Time Compute:** The process, during inference, whereby an AI model explores and evaluates different possible outputs, often in a hierarchical way following a pre-set plan, in order to select the optimal outcome of a certain task. This technique mimics human-like reasoning and deliberation.
- **Instruction Fine-tuning:** The process of further training a pre-trained model, typically a Large Language Model, to follow instructions and generate desired outputs. This involves creating datasets with diverse instructions and corresponding outputs, and training the model to understand and perform such instructions reliably.
- **JSON (JavaScript Object Notation):** A lightweight data-interchange format commonly used for transmitting data in web applications and other scenarios. It uses human-readable text.
- **LLM (Large Language Model):** A type of AI model, typically trained using deep learning techniques on vast amounts of text data. LLMs are capable of understanding, generating, and manipulating human language, and can be used for text completion, translation, summarization, and other NLP tasks.
- **Machine Learning (ML):** A subfield of AI where computer systems learn from data without being explicitly programmed. Machine learning models improve performance through exposure to data, often by adjusting their internal parameters.

- **MLOps (Machine Learning Operations):** A set of practices, processes, and tools for deploying, managing, and monitoring ML models in production. MLOps aims to streamline the entire lifecycle of ML applications from development to deployment and maintenance.
- **Modality:** A specific type of data or sensory input, such as text, images, audio, video, or sensor data.
- **Multimodal AI:** AI systems that process and integrate information from multiple modalities, allowing them to understand complex data with richer context than single-modality models. For example, an AI system that uses both images and text to better interpret a scene.
- **One-Shot Learning:** A machine learning approach where a model learns from just one labeled example per new class. This is often used in scenarios where data is extremely scarce.
- **Open Weights Model:** An AI model for which the trained parameters (or “weights”) are publicly available. This allows for greater transparency, customization, and scrutiny, but also creates challenges in terms of moderation and monitoring.
- **Overfitting:** A situation where a machine learning model learns the training data too well, including noise and random variations, resulting in poor performance on new, unseen data.
- **Pre-training:** The process of training a model on a massive dataset to create a foundational model capable of understanding various features of the data, whether it is language, images, or other forms of data. The resulting model can be further fine-tuned for specific tasks.
- **Prompt:** The input or instruction given to a language model to elicit a desired response. A well-crafted prompt is often the key to getting useful output from the model.
- **Prompt Engineering:** The process of designing and crafting effective prompts for language models to elicit the desired behavior, including task descriptions and examples. It’s a new field focused on how best to interact with AI to maximize its output quality.
- **Reasoning:** The ability of an AI system to infer, draw conclusions, and solve complex problems based on the information it has been given. This includes using techniques like chain-of-thought to work through steps to reach a logical conclusion, often with the help of inference-time compute.
- **Regression:** A supervised machine learning task where a model is trained to predict a continuous output, for example, predicting house prices or future sales numbers.
- **RNN (Recurrent Neural Network):** A type of neural network designed to process sequential data, such as time series, text, and speech. RNNs have internal feedback loops that allow them to maintain a memory of past inputs, making them suitable for data where the order matters.
- **Rule-Based System:** An early form of AI that relies on predefined rules and logic to make decisions, rather than learning from data. These systems are limited in their adaptability to new situations.
- **Semi-Structured Data:** Data that has some organizational properties but doesn’t conform to the rigid structure of relational databases. Examples include JSON and XML files, which have tags or keys but no fixed schema.
- **Supervised Learning:** A machine learning approach where models are trained on labeled datasets, in which each input has an associated correct output (label).
- **Synthetic Data:** Artificial data generated by AI models. This is used for training models where real data is insufficient, expensive, or sensitive and used to augment limited datasets, improve model fairness, or protect privacy.
- **Tensor Processing Unit (TPU):** A custom AI accelerator developed by Google, optimized for the intensive computations of deep learning models. TPUs are designed for both training and inference.
- **Tokenization:** The process of breaking down text or other inputs into smaller units called tokens, which are then used by the language model. Tokens can be words, sub-words, or characters, depending on the model and task.



- **Transfer Learning:** An approach that leverages pre-trained models to adapt knowledge to new, related tasks, reducing the need for extensive task-specific training data and computational resources. Transfer learning is often combined with fine-tuning using task-specific data to achieve better performance.
- **Transformer:** An advanced type of neural network architecture that uses attention mechanisms to process and generate data. Transformers are particularly effective for sequence-to-sequence tasks, such as natural language translation, text generation, and code completion. They are the foundation for many modern large language models.
- **Unsupervised Learning:** A machine learning approach where models learn from unlabeled data by identifying patterns and structures within the data, without specific output labels.
- **XML (Extensible Markup Language):** A markup language used for encoding documents in a way that is both human-readable and machine-readable. Often used for transmitting data between systems, especially in web services.
- **Zero-Shot Learning:** An AI technique that allows models to perform tasks they were not explicitly trained on, i.e. allowing models to generalize to classes that they haven't seen during training by leveraging auxiliary information or understanding relationships between categories or classes, for example, identifying a new animal species based on its description without prior exposure to any images of the new species.

References

- Aggarwal, C. C., & Reddy, C. K. (2014). *Data Clustering: Algorithms and Applications*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Chapman & Hall/CRC.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877–1901.
- Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *In Proceedings of the 23rd international conference on Machine learning*, 161–168.
- Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3), 1–45.
- Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. (2021). Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374*.
- Cottier, B., Rahman, R., Fattorini, L., Maslej, N., & Owen, D. (2024). The rising costs of training frontier ai models.
- Devlin, J., Chang, M-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78–87.
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
- Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., & Narasimhan, K. (2024). SWE-bench: Can Language Models Resolve Real-World GitHub Issues? *arXiv preprint arXiv:2310.06770*.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
- Hoffer, E., Hubara, I., & Soudry, D. (2017). Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in neural information processing systems*, 30.
- Irugalbandara, C., Mahendra, A., Daynauth, R., Arachchige, T. K., Dantanarayana, J., Flautner, K., Tang, L., Kang, Y., & Mars, J. (2024). Scaling down to scale up: A cost-benefit analysis of replacing openai's llm with open source slms in production. *In 2024 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 280–291. IEEE.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), 1–35.
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 2053951716679679.
- OpenAI et al. (2024). OpenAI o1 System Card. *arXiv preprint arXiv:2412.16720*.
- Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10), 1345–1359.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 9.
- Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-Shot Text-to-Image Generation. *In Proceedings of the 38th International Conference on Machine Learning*, 139, 8821–8831.



- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?" explaining the predictions of any classifier. *In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.
- Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Pearson.
- Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in neural information processing systems*, 28.
- Squicciarini, M., Genao, J. V., & Sarmiento, C. (2024). Synthetic content and its implications for ai policy: A primer. Paris: UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000392181>
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56), 1929–1958.
- Vaswani, A., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3), 1–34.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint arXiv:2201.11903*.
- Xian, Y., Lampert, C. H., Schiele, B., & Akata, Z. (2018). Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9), 2251–2265.

Artificial Intelligence and the Evolution of AI (Model) Capabilities

A Conceptual Primer

The excitement around AI is leading citizens and policy makers alike to express fears and hopes about something that, nevertheless, few know well. Too often, many talk about AI without having an actual understanding about what AI is and is not, what it can do or is unable to, and how AI systems actually work — their capabilities, limitations, and the ways they learn and make decisions.

This primer addresses the need for a clear understanding of AI technologies, how they have evolved from more basic rule-based systems to generative AI and increasingly complex agentic reasoning systems. Such an understanding is paramount for the design and implementation of effective and adaptive governance frameworks, as those advocated in UNESCO Recommendation on the Ethics of Artificial Intelligence.



unesco

Social and Human Sciences Sector

7, place de Fontenoy
75352 Paris 07 SP France

✉ ai-ethics@unesco.org

🌐 on.unesco.org/EthicsAI

Follow us

@UNESCO #AI #HumanAI

