LARGE LANGUAGE MODELS, DEEP LEARNING, AND MACHINE LEARNING: A COMPREHENSIVE REVIEW OF ARCHITECTURES, APPLICATIONS, CHALLENGES, AND FUTURE DIRECTIONS

Akash Saraswat[1], Dr. Dinesh Chandra Misra[2], Dr. Rahul Kumar[3]

PhD Scholar[1], Associate Professor[2,3]

[1,2]Department of Computer Science and Engineering, Dr. K. N. Modi University

[3]Department Of Computing and Informatics, Sir Padampat Singhania University, Udaipur Rajasthan

sharma9910483702@gmail.com[1], dcmishra99@gmail.com[2] , rahulkumar1680@gmail.com[3]

Abstract

The rapid advancement of artificial intelligence (AI) has transformed the development of intelligent systems, with Machine Learning (ML), Deep Learning (DL), and Large Language Models (LLMs) emerging as major technological paradigms. This review provides a comprehensive analysis of these approaches, focusing on their architectures, learning mechanisms, applications, advantages, limitations, and future research directions. Traditional ML techniques, including Support Vector Machines, Decision Trees, Random Forests, and Gradient Boosting, have demonstrated strong performance in structured-data prediction and classification tasks but often depend on manually engineered features. DL addresses these limitations through multilayer neural architectures capable of automatically learning hierarchical representations from complex data. Convolutional Neural Networks, Recurrent Neural Networks, Long Short-Term Memory networks, and Transformer-based architectures have consequently achieved significant advances in image, speech, video, and language processing. Building upon Transformer architectures, LLMs have introduced a new generation of AI systems capable of performing language understanding, generation, summarization, question answering, reasoning, code generation, and multimodal tasks. This review examines important LLM development strategies, including pre-training, fine-tuning, instruction tuning, reinforcement learning from human feedback, parameter-efficient fine-tuning, prompt engineering, and Retrieval-

Augmented Generation. Applications across healthcare, education, finance, cybersecurity, agriculture, business, and software engineering are also examined. Despite their capabilities, ML, DL, and LLM-based systems face challenges related to data quality, computational requirements, interpretability, model bias, privacy, security, robustness, hallucination, and energy consumption. The review further identifies emerging research opportunities involving explainable AI, efficient and smaller language models, multimodal learning, domain-specific LLMs, continual learning, federated learning, edge intelligence, and hybrid ML-DL-LLM frameworks. Overall, the study highlights the evolutionary relationship among ML, DL, and LLMs and emphasizes that their integration can provide more scalable, adaptive, reliable, and intelligent solutions for next-generation AI applications.

Keywords: Large Language Models, Machine Learning, Deep Learning, Artificial Intelligence, Transformer Architecture, Natural Language Processing, Neural Networks, Generative AI, Retrieval-Augmented Generation, Explainable AI, Multimodal Learning, Intelligent Systems

1. Introduction

Artificial Intelligence (AI) refers to the broad field of computer science concerned with developing computational systems capable of performing tasks that normally require human intelligence, including learning, reasoning, problem-solving, perception, language understanding, planning and decision-making. The development of AI has progressed from rule-based systems and symbolic reasoning toward data-driven approaches that learn patterns from examples. Modern AI increasingly relies on statistical learning, neural networks and large-scale computational infrastructure to develop systems that can adapt to complex environments. AI is now applied across healthcare, finance, education, transportation, manufacturing, cybersecurity, scientific research and communication. The growing availability of large datasets, powerful graphical processing units (GPUs), cloud computing and advanced optimization methods has accelerated this transformation. Rather than representing a single algorithm or technology, AI encompasses multiple approaches, including machine learning, deep learning, natural language processing, computer vision, knowledge representation and intelligent agents (Russell and Norvig, 2021).

Machine Learning (ML) is a major subfield of AI in which computational models learn patterns and relationships from data instead of being explicitly programmed with rules for every individual task. Machine learning algorithms can generally be categorized into supervised, unsupervised and reinforcement learning. Supervised learning uses labelled examples to perform tasks such as classification and regression, while unsupervised learning identifies hidden structures or patterns within unlabelled data. Reinforcement learning enables an agent to learn through interactions with an environment using rewards and penalties. Common ML algorithms include linear regression, logistic regression, decision trees, random forests, support vector machines, k-nearest neighbours, naïve Bayes and gradient-boosting methods. Traditional ML has been particularly effective for structured datasets where meaningful features can be identified and represented before model training. However, its performance can depend considerably on the quality of feature engineering, data preprocessing and domain knowledge (Mitchell, 1997; Bishop, 2006).

Deep Learning (DL) represents a specialized area of machine learning based on artificial neural networks containing multiple computational layers. Unlike many traditional ML methods, deep learning models can automatically learn increasingly complex representations directly from raw or minimally processed data. Early layers may learn relatively simple patterns, while deeper layers combine these representations into more abstract features. Convolutional Neural Networks (CNNs), for example, have achieved significant success in image and visual recognition, while Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) have been widely applied to sequential and temporal data. The development of deeper architectures, improved optimization algorithms, large datasets and GPU-based computation has contributed substantially to the rapid development of deep learning. Deep learning has therefore become fundamental to modern computer vision, speech recognition, recommendation systems and natural language processing (LeCun, Bengio and Hinton, 2015).

Natural Language Processing (NLP) is an interdisciplinary area of AI concerned with enabling computers to process, understand, interpret and generate human language.

Human language presents substantial computational challenges because meaning depends on context, syntax, semantics, ambiguity, cultural factors and relationships between words. Earlier NLP systems frequently relied on manually constructed linguistic rules and statistical models, whereas modern NLP increasingly uses neural networks and Transformer-based architectures. NLP applications include machine translation, sentiment analysis, text classification, information extraction, speech-to-text processing, question answering, summarization, dialogue systems and text generation. The introduction of distributed word representations and neural language models significantly improved the ability of machines to capture relationships between linguistic units. Transformer-based models subsequently enabled large-scale contextual language representations and provided the technological foundation for many contemporary NLP systems (Manning and Schütze, 1999; Goldberg, 2017).

Transformers represent a major architectural development in deep learning and have fundamentally changed modern NLP. The Transformer architecture was introduced by Vaswani et al. (2017) as an alternative to recurrent approaches for sequence processing. Its central mechanism is self-attention, which enables a model to evaluate relationships between different elements of an input sequence and determine which elements are most relevant to one another. Unlike recurrent architectures that process sequences sequentially, Transformers can process many sequence elements in parallel during training, making them particularly suitable for large-scale computation. The original architecture consisted of encoder and decoder components and incorporated multi-head attention, positional representations and feed-forward neural networks. Transformer architectures subsequently became the foundation for models such as BERT, GPT and T5, demonstrating that attention-based architectures could be effectively scaled to increasingly large datasets and model parameters (Vaswani et al., 2017; Devlin et al., 2019).

Large Language Models (LLMs) are large-scale deep learning models trained on extensive collections of textual data to learn statistical and semantic relationships within language. Most contemporary LLMs are based on Transformer architectures, particularly decoder-based architectures for autoregressive text generation. During pre-training, an LLM learns from large quantities of text through self-supervised

objectives, allowing it to acquire linguistic patterns, contextual associations and various forms of knowledge represented within its training data. Following pre-training, models can be adapted through supervised fine-tuning, instruction tuning, reinforcement learning from human feedback and parameter-efficient fine-tuning. Modern LLMs can perform a broad range of tasks, including text generation, summarization, translation, question answering, code generation, information extraction and conversational interaction. Models such as GPT and other large Transformer-based systems have demonstrated that scaling model size, training data and computational resources can produce increasingly general-purpose language capabilities (Brown et al., 2020; Bommasani et al., 2021).

The relationship between ML, DL and LLMs can be understood as a progression of increasingly specialized and powerful learning paradigms. Artificial Intelligence is the broadest field, while machine learning constitutes one of its major approaches. Deep learning is a subset of machine learning that uses multilayer neural networks to learn representations from data. Transformers are a particular deep learning architecture based primarily on attention mechanisms, and LLMs are large-scale language models that predominantly employ Transformer architectures. Therefore, the conceptual progression can be represented as **Artificial Intelligence → Machine Learning → Deep Learning → Transformers → Large Language Models**. However, this progression should not be interpreted as a complete replacement of one technology by another. Traditional ML remains highly valuable for structured data and resource-constrained applications, while DL is particularly effective for complex unstructured data. LLMs extend deep learning capabilities primarily into language-intensive and increasingly multimodal applications. Hybrid systems can consequently combine traditional ML predictors, deep learning feature extractors and LLM-based reasoning or interaction components (LeCun, Bengio and Hinton, 2015; Vaswani et al., 2017; Bommasani et al., 2021).

Traditional machine learning relies substantially on feature engineering, in which domain-specific characteristics are selected, transformed or constructed before the learning algorithm is applied. For example, a predictive system may require researchers to identify meaningful numerical, statistical or categorical features from

raw observations. Deep learning reduces this dependence by learning hierarchical representations automatically. This difference is particularly important when dealing with unstructured data such as images, audio and natural language, where manually defining all relevant features can be difficult. Deep neural networks can discover multiple levels of representation through training, allowing the model to identify complex patterns that may not be easily captured through manually designed features (LeCun, Bengio and Hinton, 2015). LLMs take this principle further by learning language representations from extremely large datasets and using these representations across multiple downstream tasks. Consequently, the transition from traditional ML to DL and then to LLMs reflects not simply an increase in model size but a broader shift toward automated representation learning, large-scale pre-training and increasingly general-purpose AI systems.

The motivation for this review arises from the rapid convergence of machine learning, deep learning, NLP and large language models within contemporary AI research. Although these technologies are closely connected, they differ considerably in terms of architecture, data requirements, computational cost, training strategies, interpretability and application domains. The rapid emergence of LLMs has also created a need to understand them within the wider historical and technical development of machine learning and deep learning rather than considering them as isolated technologies. Existing research frequently focuses on specific aspects such as deep learning architectures, Transformer models, LLM training or individual applications. A comprehensive perspective that connects these developments can provide a clearer understanding of how modern intelligent systems have evolved and why particular architectures are appropriate for different problems. The increasing deployment of generative and language-based AI in real-world environments further strengthens the need to critically examine their reliability, safety, scalability and practical limitations (Bommasani et al., 2021).

Several research gaps remain despite the substantial progress in these fields. One important gap concerns the reliability and interpretability of increasingly complex AI models. Traditional ML models can sometimes provide relatively understandable decision processes, whereas deep neural networks and LLMs can contain millions or

billions of parameters, making their internal reasoning difficult to interpret. LLMs also face challenges related to hallucination, bias, misinformation, robustness and sensitivity to prompts. Another important gap concerns computational efficiency because training and deploying large models can require substantial computational resources, energy and financial investment. Research is therefore increasingly required on smaller language models, parameter-efficient adaptation, model compression, quantization and efficient inference. Additional gaps exist in privacy-preserving learning, domain-specific adaptation, continual learning, multilingual performance, low-resource languages, multimodal reasoning and standardized evaluation. These challenges demonstrate that achieving high predictive or generative performance alone is insufficient; future AI systems must also be trustworthy, efficient, transparent and robust (Bender et al., 2021; Bommasani et al., 2021).

This review contributes to the existing body of knowledge by presenting an integrated examination of machine learning, deep learning and large language models within the broader development of artificial intelligence. It establishes a conceptual connection between traditional learning algorithms, neural representation learning, Transformer architectures and contemporary LLMs. The review further compares their fundamental architectures, learning mechanisms, data requirements, computational characteristics and application areas while examining their respective strengths and limitations. Particular attention is given to modern LLM development strategies, including pre-training, fine-tuning, instruction tuning, reinforcement learning from human feedback and retrieval-based approaches. In addition, the review identifies important challenges involving hallucination, bias, explainability, computational efficiency, privacy and security and discusses emerging directions such as multimodal models, efficient LLMs, Retrieval-Augmented Generation, continual learning, federated learning and hybrid AI systems. Through this integrated perspective, the review aims to provide researchers and practitioners with a structured understanding of the evolution of intelligent systems and the opportunities for developing more reliable, efficient and generalizable AI technologies.

2. Evolution of Machine Learning to LLMs

The evolution from Machine Learning (ML) to Large Language Models (LLMs) represents a major transformation in the development of artificial intelligence. This progression has moved from systems that depend heavily on manually designed features and task-specific algorithms toward highly scalable neural architectures capable of learning representations directly from massive datasets. Machine learning initially emerged as a practical approach for enabling computers to learn patterns from data without requiring explicit programming for every decision. Early ML methods included linear regression, logistic regression, decision trees, support vector machines, naïve Bayes and k-nearest neighbours. These approaches achieved considerable success in classification, regression and pattern recognition, particularly when datasets were structured and informative features could be identified by domain experts. However, their effectiveness often depended on feature engineering, preprocessing and careful selection of algorithms, creating limitations when dealing with highly complex and unstructured data (Mitchell, 1997; Bishop, 2006).

The development of Deep Learning (DL) addressed several of these limitations by introducing multilayer artificial neural networks capable of automatically learning representations from data. Instead of relying entirely on manually engineered features, deep learning models learn multiple levels of abstraction during training. Convolutional Neural Networks became highly effective for image and visual recognition, while Recurrent Neural Networks and Long Short-Term Memory networks were widely applied to sequential information such as speech and text. The combination of increasing computational power, large datasets and improved optimization techniques significantly accelerated the adoption of deep learning. The success of deep neural networks demonstrated that representation learning could outperform conventional feature-based approaches for many complex tasks, particularly in computer vision, speech recognition and natural language processing (LeCun, Bengio and Hinton, 2015).

Within Natural Language Processing (NLP), the transition from conventional statistical approaches to neural language models represented another important stage

in this evolution. Traditional NLP systems commonly depended on manually developed linguistic rules, statistical representations and features such as word frequencies and n-grams. Neural approaches introduced distributed representations in which words and other linguistic units could be represented as vectors within continuous mathematical spaces. These representations enabled models to capture semantic and syntactic relationships more effectively than many traditional symbolic or statistical methods. Recurrent neural architectures subsequently improved the processing of sequential language by allowing information from previous words to influence the interpretation of later words. Nevertheless, recurrent models faced difficulties with long-range dependencies and sequential computation, which limited their scalability for very large datasets (Goldberg, 2017).

A major turning point occurred with the introduction of the Transformer architecture by Vaswani et al. (2017). Transformers replaced recurrent processing with self-attention mechanisms that allow models to identify relationships between different elements of a sequence while processing them more efficiently in parallel. This architecture provided a scalable foundation for large-scale language modelling and subsequently became central to modern NLP. Transformer-based models such as BERT demonstrated the effectiveness of large-scale pre-training for language understanding, while autoregressive models such as GPT demonstrated the ability of large neural networks to generate coherent and contextually relevant text (Devlin et al., 2019; Brown et al., 2020).

The emergence of LLMs represents the latest major stage in this progression. LLMs are generally Transformer-based models trained on extremely large collections of textual data using self-supervised learning objectives. Instead of being developed exclusively for one specific task, these models can acquire broad language representations during pre-training and subsequently perform multiple tasks through prompting, fine-tuning or instruction-based adaptation. Scaling model parameters, training datasets and computational resources has contributed to increasingly capable language systems capable of summarization, translation, question answering, dialogue, code generation and other complex language tasks (Brown et al., 2020; Bommasani et al., 2021). Consequently, the evolution can be represented as **Machine**

Learning → Deep Learning → Neural NLP → Transformers → Large Language Models, reflecting a broader movement toward automated representation learning, large-scale pre-training and increasingly general-purpose AI systems.

3. Evolution of Deep Learning to LLMs

The evolution of Deep Learning (DL) to Large Language Models (LLMs) represents one of the most significant developments in modern artificial intelligence. Deep learning emerged as an advanced form of machine learning based on artificial neural networks with multiple interconnected layers that can automatically learn hierarchical representations from large and complex datasets. Unlike traditional machine learning methods, which often require extensive manual feature engineering, deep learning models learn relevant features directly from raw data during the training process. The availability of large datasets, graphical processing units (GPUs), improved optimization techniques and advances in neural network architectures contributed substantially to the rapid development of deep learning. These developments enabled neural networks to achieve remarkable performance in image classification, speech recognition, natural language processing and other complex tasks (LeCun, Bengio and Hinton, 2015).

Early deep learning architectures included feed-forward neural networks, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). CNNs became particularly successful in computer vision because they could automatically identify spatial features such as edges, textures and shapes from images. RNNs, on the other hand, were designed to process sequential information and became important for applications involving text, speech and time-series data. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks were subsequently introduced to address some of the limitations of conventional RNNs, particularly difficulties in learning long-term dependencies. These architectures significantly improved sequence modelling and were widely used in machine translation, speech recognition and language modelling. However, recurrent networks processed sequences sequentially, making them computationally expensive and

difficult to scale efficiently for increasingly large datasets and models (Hochreiter and Schmidhuber, 1997; Cho et al., 2014).

The development of neural language models created an important connection between deep learning and Natural Language Processing (NLP). Earlier NLP systems relied heavily on manually defined linguistic rules and statistical representations, whereas neural approaches enabled language information to be represented through continuous vector representations. Word embedding techniques such as Word2Vec allowed words to be represented as dense numerical vectors in which semantic and syntactic relationships could be captured. Neural language models subsequently became increasingly capable of understanding contextual relationships between words and sentences. Nevertheless, conventional embeddings generally assigned similar representations to a word regardless of its specific context, creating difficulties for words with multiple meanings. These limitations encouraged researchers to develop contextual language representations and larger neural architectures (Mikolov et al., 2013).

A major breakthrough occurred with the introduction of the Transformer architecture. Vaswani et al. (2017) proposed the Transformer as an architecture based primarily on self-attention rather than recurrent processing. Self-attention enables the model to examine relationships between different elements of a sequence and determine which elements are important for interpreting a particular word or token. Multi-head attention allows the model to learn different types of relationships simultaneously, while positional representations provide information about the order of tokens. Importantly, Transformers allow much greater parallelization during training than recurrent architectures, making them highly suitable for large-scale computation. This scalability became a fundamental factor in the development of increasingly large language models (Vaswani et al., 2017).

Transformer-based pre-training subsequently transformed NLP research. BERT introduced a bidirectional Transformer-based approach that learned contextual language representations through large-scale pre-training and could then be fine-tuned for specific downstream tasks (Devlin et al., 2019). In parallel, the GPT family

demonstrated the effectiveness of autoregressive Transformer architectures for language generation. GPT-3 showed that scaling the number of parameters and training data could produce strong performance across multiple tasks, including translation, question answering and text generation, often with only a small number of task-specific examples (Brown et al., 2020). These developments established the principle that a single large pretrained model could serve as the foundation for numerous language-based applications.

Large Language Models represent the next stage of this deep learning evolution. LLMs are generally based on Transformer architectures and are trained using enormous quantities of textual data and computational resources. During pre-training, the models learn statistical relationships and contextual patterns within language through self-supervised learning. They can subsequently be adapted through fine-tuning, instruction tuning, prompt engineering, parameter-efficient fine-tuning and reinforcement learning techniques. Modern LLMs can perform diverse tasks such as text generation, summarization, translation, question answering, code generation, information extraction and conversational interaction (Bommasani et al., 2021). The emergence of LLMs therefore demonstrates how deep learning has progressed from task-specific neural architectures toward highly scalable and general-purpose models capable of supporting a wide range of intelligent applications.

Overall, the evolution from deep learning to LLMs can be understood as a progression from **multilayer neural networks → CNNs and RNNs → LSTM/GRU networks → neural language models → Transformers → pretrained language models → Large Language Models**. Each stage addressed limitations of previous approaches while increasing the ability of models to learn complex representations from data. The transition to Transformer-based LLMs has particularly emphasized large-scale pre-training, model scaling and transferability across tasks. However, this progression has also introduced challenges involving computational cost, model interpretability, bias, privacy, hallucination, robustness and energy consumption. Consequently, current research increasingly focuses not only on developing larger models but also on creating efficient, trustworthy, explainable and domain-adaptable LLM systems (Bender et al., 2021; Bommasani et al., 2021).

4. Comparison of ML, DL and LLM

Machine Learning (ML), Deep Learning (DL), and Large Language Models (LLMs) represent interconnected stages in the development of modern artificial intelligence, but they differ considerably in their learning mechanisms, architectures, data requirements, computational complexity and application domains. Machine learning is a broad approach in which algorithms learn patterns from data and use those patterns to make predictions or decisions. Traditional ML methods, such as Support Vector Machines, Decision Trees, Random Forests, Logistic Regression and Gradient Boosting, are particularly effective for structured datasets. These algorithms commonly require data preprocessing and feature engineering, where relevant characteristics are manually selected or transformed before training. Consequently, the quality of the final model can depend strongly on the expertise used to construct the input features (Mitchell, 1997; Bishop, 2006).

Deep Learning is a specialized branch of machine learning that uses multilayer neural networks to automatically learn representations from data. Instead of depending primarily on manually engineered features, DL models learn increasingly complex features through multiple computational layers. For example, a CNN can learn edges and textures in early layers and progressively combine them into higher-level visual patterns. Similarly, recurrent architectures such as LSTM networks can learn temporal dependencies in sequential data. Deep learning has demonstrated exceptional performance in computer vision, speech recognition and NLP, particularly when sufficient training data and computational resources are available (LeCun, Bengio and Hinton, 2015). Compared with conventional ML, DL generally requires more data and computational power but can achieve superior performance on complex unstructured datasets.

LLMs represent a specialized category of large-scale deep learning systems designed primarily for processing and generating natural language. Most modern LLMs are based on Transformer architectures, which use self-attention mechanisms to model relationships among tokens within sequences (Vaswani et al., 2017). Unlike traditional ML models that are generally trained for a particular task and many

conventional DL systems that require task-specific training, LLMs can be pretrained on massive datasets and subsequently adapted to numerous tasks through prompting, fine-tuning or instruction tuning. Models such as GPT-3 demonstrated that sufficiently large pretrained language models could perform multiple NLP tasks with limited task-specific examples (Brown et al., 2020). This general-purpose capability represents one of the most important distinctions between LLMs and conventional ML or DL approaches.

Table 1. Comparative Characteristics of Machine Learning, Deep Learning, and Large Language Models

Characteristic	**Machine Learning (ML)**	**Deep Learning (DL)**	**Large Language Models (LLMs)**
Basic principle	Learns patterns using statistical algorithms	Learns hierarchical representations using neural networks	Learns large-scale language representations using deep neural networks
Feature engineering	Often required	Mostly automated	Primarily learned automatically
Typical architecture	Trees, SVM, regression, boosting	CNN, RNN, LSTM, GRU, Transformer	Primarily Transformer-based
Data requirement	Low to moderate	High	Very high
Computational requirement	Low to moderate	High	Very high
Training approach	Usually task-specific	Task-specific or pretrained	Large-scale pre-training and adaptation
Main data type	Structured data	Images, audio, text, signals	Text and increasingly multimodal data
Generalization	Usually limited to related tasks	Better transfer capabilities	Strong multitask and transfer capabilities
Interpretability	Relatively higher	Generally lower	Often highly difficult

Typical applications	Classification, regression, prediction	Vision, speech, NLP, forecasting	Generation, dialogue, reasoning, summarization, coding

Another major difference concerns the scale and nature of training. Traditional ML models can often achieve useful performance with relatively small or medium-sized datasets because their architectures are less parameter-intensive. Deep learning models generally benefit from larger datasets because millions of parameters may need to be optimized during training. LLMs take this scaling principle much further by using enormous datasets and billions of parameters in some models. Large-scale pre-training enables LLMs to develop general-purpose representations that can subsequently be reused for different tasks, reducing the need to train an independent model for every application (Bommasani et al., 2021). However, this advantage comes with substantially greater requirements for computational infrastructure, memory, energy and training time.

The three approaches also differ in their applications. Traditional ML remains highly valuable for structured prediction problems, including credit-risk assessment, fraud detection, customer classification and tabular medical prediction. DL is particularly suitable for complex perceptual tasks, including medical image classification, object detection, speech recognition and time-series analysis. LLMs are primarily associated with language-intensive applications such as conversational agents, document summarization, question answering, translation, content generation and code generation. Nevertheless, the boundaries between these categories are becoming increasingly flexible. Modern intelligent systems can combine ML models for structured prediction, DL models for feature extraction and LLMs for natural-language interaction and reasoning.

Table 2. Comparative Analysis of Learning, Computational, and Application Aspects of ML, DL, and LLMs

Aspect	ML	DL	LLM
Learning representation	Manually engineered + learned	Automatically learned	Large-scale contextual representations
Model scale	Small–large	Large	Very large
Training cost	Relatively low	High	Extremely high
Inference cost	Usually low	Moderate–high	Moderate–very high
Task specialization	High	Moderate–high	Relatively low
Language capability	Limited	Strong with appropriate architecture	Advanced
Multimodal potential	Limited	High	Increasingly high
Adaptation	Retraining/feature modification	Fine-tuning	Prompting, fine-tuning, PEFT, RAG
Key limitation	Feature dependence	Data and computation	Hallucination, cost, bias and interpretability

Despite these differences, ML, DL and LLMs should not be viewed as competing technologies. Rather, they form complementary components of the broader AI ecosystem. Deep learning itself is a subset of machine learning, while LLMs represent a highly specialized application of large-scale deep learning, particularly through Transformer architectures. The progression from ML to DL and LLMs reflects increasing automation of feature learning, model scale and generalization capabilities. However, traditional ML can remain more efficient and interpretable for many structured-data applications, while smaller DL models may be preferable where computational resources are constrained. LLMs provide powerful language capabilities but introduce challenges related to hallucination, bias, security, privacy and computational requirements (Bender et al., 2021). Therefore, future AI systems

are likely to increasingly adopt hybrid ML-DL-LLM architectures, combining the efficiency of conventional ML, representation-learning capabilities of DL and language-based reasoning and interaction capabilities of LLMs.

5. Conclusion and Future Work

Machine Learning (ML), Deep Learning (DL), and Large Language Models (LLMs) represent important stages in the continuous evolution of artificial intelligence. ML established the foundation for data-driven prediction and decision-making by enabling computational algorithms to identify patterns without explicit programming. Deep learning extended this capability through multilayer neural networks that automatically learn hierarchical representations from complex and unstructured data. The development of Transformer architectures subsequently transformed natural language processing by introducing self-attention mechanisms capable of efficiently modelling relationships across long sequences. Building upon these developments, LLMs have emerged as highly scalable deep learning systems capable of performing diverse language-related tasks, including text generation, summarization, translation, question answering, code generation and conversational interaction (LeCun, Bengio and Hinton, 2015; Vaswani et al., 2017; Brown et al., 2020). The evolution from ML to DL and LLMs therefore demonstrates a broader movement from manually engineered features and task-specific models toward automated representation learning, large-scale pre-training and increasingly general-purpose intelligent systems.

The comparative analysis demonstrates that each paradigm offers distinct advantages and limitations. Traditional ML remains highly effective for structured datasets, relatively small-scale applications and situations where computational efficiency and interpretability are important. DL provides superior capabilities for learning complex representations from images, audio, text and other high-dimensional data, although it generally requires greater quantities of training data and computational resources. LLMs extend deep learning toward general-purpose language intelligence and can perform multiple tasks through prompting, fine-tuning and instruction-based adaptation. However, their increasing scale introduces significant challenges, including high computational costs, energy consumption, limited interpretability, bias,

privacy concerns, security vulnerabilities and hallucination, where models generate plausible but factually incorrect information (Bender et al., 2021; Bommasani et al., 2021). Therefore, model performance alone cannot be considered sufficient for evaluating next-generation AI systems; reliability, transparency, safety, efficiency and responsible deployment are equally important.

Future research should focus on developing more efficient, reliable and adaptable AI systems rather than concentrating exclusively on increasing model size. Parameter-efficient fine-tuning, quantization, pruning, knowledge distillation and smaller language models can reduce the computational and financial requirements associated with LLM deployment. Retrieval-Augmented Generation (RAG) and domain-specific adaptation can further improve factuality by allowing models to access relevant and up-to-date external information. Research into explainable AI should also continue to improve understanding of how deep learning and LLM systems reach their outputs, particularly in high-stakes domains such as healthcare, finance and legal services. In addition, continual learning is an important direction because future AI systems will need to incorporate new information while retaining previously acquired knowledge without substantial performance degradation.

Another promising direction is the development of multimodal and hybrid AI architectures. Future systems may integrate text, images, audio, video, sensor data and structured information within a unified framework. Hybrid ML-DL-LLM systems could combine the efficiency of conventional machine learning, the representation-learning capabilities of deep learning and the language reasoning and interaction capabilities of LLMs. Federated learning and privacy-preserving techniques may provide opportunities for deploying AI in environments where sensitive data cannot be centrally collected. Greater attention is also required for multilingual and low-resource language models to reduce technological disparities between languages and regions. Finally, standardized evaluation frameworks should assess not only accuracy and language quality but also robustness, fairness, factuality, hallucination, safety, energy efficiency and real-world reliability. Collectively, these research directions can support the transition from highly capable AI models toward **efficient, trustworthy, explainable, sustainable and human-centred intelligent systems**.

References

1. Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S. (2021) 'On the dangers of stochastic parrots: Can language models be too big?', *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623.
2. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N.S., Chen, A., Creel, K., Davis, J.Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D.E., Hong, J., Hsu, K., Huang, J., Ibarz, B., Ichter, B., Ilyas, A., Jain, S., Jia, R., Kannan, S., Kaplan, J., Khani, F., Kim, D., Lee, T., Leike, J., Lejay, A., K. et al. (2021) 'On the opportunities and risks of foundation models', *arXiv preprint arXiv:2108.07258*.
3. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. and Amodei, D. (2020) 'Language models are few-shot learners', *Advances in Neural Information Processing Systems*, 33, pp. 1877–1901.
4. Bishop, C.M. (2006) *Pattern Recognition and Machine Learning*. New York: Springer.
5. Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019) 'BERT: Pre-training of deep bidirectional transformers for language understanding', *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1, pp. 4171–4186.
6. Goldberg, Y. (2017) *Neural Network Methods for Natural Language Processing*. San Rafael, CA: Morgan & Claypool Publishers.
7. Hochreiter, S. and Schmidhuber, J. (1997) 'Long short-term memory', *Neural Computation*, 9(8), pp. 1735–1780.

8. LeCun, Y., Bengio, Y. and Hinton, G. (2015) 'Deep learning', *Nature*, 521, pp. 436–444.
9. Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) 'Efficient estimation of word representations in vector space', *arXiv preprint arXiv:1301.3781*.
10. Mitchell, T.M. (1997) *Machine Learning*. New York: McGraw-Hill.
11. Russell, S. and Norvig, P. (2021) *Artificial Intelligence: A Modern Approach*. 4th edn. Hoboken, NJ: Pearson.
12. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) 'Attention is all you need', *Advances in Neural Information Processing Systems*, 30, pp. 5998–6008.