Software: Practice and Experience

WILEY

RESEARCH ARTICLE

Large AI Models and Their Applications: Classification, Limitations, and Potential Solutions

Jing Bi[1] | Ziqi Wang[1] | Haitao Yuan[2] | Xiankun Shi[1] | Ziyue Wang[3] | Jia Zhang[4] | MengChu Zhou[5] | Rajkumar Buyya[6]

[1]College of Computer Science, Beijing University of Technology, Beijing, China | [2]School of Automation Science and Electrical Engineering, Beihang University, Beijing, China | [3]Mechanical Electrical Engineering School, Beijing Information Science and Technology University, Beijing, China | [4]Department of Computer Science in the Lyle School of Engineering, Southern Methodist University, Dallas, USA | [5]Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, USA | [6]Cloud Computing and Distributed Systems (CLOUDS) Lab, School of Computing and Information Systems, The University of Melbourne, Melbourne, Australia

Correspondence: Haitao Yuan (yuan@buaa.edu.cn)

Received: 16 August 2024 | **Revised:** 22 December 2024 | **Accepted:** 3 January 2025

Funding: This work was supported by the National Natural Science Foundation of China under Grants 62173013 and 62473014, the Beijing Natural Science Foundation under Grants L233005 and 4232049, and Beihang World TOP University Cooperation Program.

Keywords: artificial intelligence | autonomous driving technologies | large models | neural network | smart industrial productions

ABSTRACT

Background: In recent years, Large Models (LMs) have been rapidly developed, including large language models, visual foundation models, and multimodal LMs. They are updated and iterated at a very fast pace. These LMs can accomplish many tasks, *e.g.*, daily work assistant, intelligent customer service, and intelligent factory scheduling. Their development has contributed to various industries in human society.

Aims: The architectural flaws of LMs lead to several problems, including illusions and difficulty in locating errors, limiting their performance. Solving these problems properly can facilitate their further development.

Methods: This work first introduces the development of LMs and identifies their current problems, including data and energy consumption, catastrophic forgetting, reasoning ability, localization fault, and ethical problems. Then, potential solutions to these problems are provided, including increase data and computation capability, neural-symbolic synergy, and data orientation to human pattern.

Discussion: This work discusses developing vertical domain LMs on top of some base LMs. In addition, this work introduces three typical real-world applications of LMs, including autonomous driving, smart industrial productions, and intelligent medical assistance.

Conclusion: By embracing the advantages of LMs and solving their fundamental problems, many industries are expected to achieve promising prospects in the future.

1 | Introduction

Large models (LMs) are neural networks that contain super-large-scale parameters. They can perform tasks in areas such as content generation that only human beings could perform in the past. They are considered to be a sign that artificial intelligence (AI) has changed from the weak to strong [1]. As noted in Figure 1, the first-generation AI in the 1970s focused

Abbreviations: AI, artificial intelligence; CLIP, contrastive language-image pretraining; FSD, full-self driving; GPT, generative pretrained transformer; LMs, large models.

https://doi.org/10.1002/spe.3408

on computational intelligence, concentrating on the computation tasks and data storage. Second-generation AI in the 2000s focused on perceptive intelligence, emphasizing the recognition and perception of tasks in different modalities. Third-generation AI in the 2020s focused on cognitive intelligence. It focuses on understanding and reflecting the external environment, which perceives the external environment and realizes decision-making and scheduling of tasks. The goal of LMs, as third-generation AI, is to make quick decisions after sensing the external environment to help human beings. Their development is also inspired by a large number of milestone discoveries and inventions. The back-propagation algorithm [2] proposed in 1984 solves the problem of training neural networks; the universal approximation theorem [3] proved in 1989 theoretically shows the powerful fitting ability of neural networks, stating that once the network has a sufficient number of neurons, it can fit any complex continuous function; Transformer [4] proposed in 2017 can capture long and short dependencies and achieve highly parallelized computation by removing the loop structure; self-supervised learning proposed in 2018 [5] addresses the problem of training on unlabeled data; neural scaling rate in 2020 [6] reveals a positive correlation among the number of parameters, amount of data, computational power, and performance, showing that the predictive performance of a model improved according to a power law with more data and a larger model. The essence of LMs is to use powerful algorithms and large amounts of computational power to train complex probability distribution functions from massive amounts of data. Over the past few years, the development of LMs experienced rapid evolution and expansion.

The development of LMs is shown in Figure 2. In the early exploratory phase in 2018, the first generation of generative pre-trained transformer (GPT-1) led to a new paradigm of natural language processing. Following this, GPT-2 in 2019 grew in size and made significant progress in understanding and generating text. From 2020 to 2021, LMs entered a period of extensive exploration. During this time, GPT-3 set a new standard with its staggering 175 billion parameters, catalyzing the emergence of a series of innovative LMs. Huawei's PanGu-α excelled in text generation fields such as knowledge question answering, retrieval, reasoning, and reading comprehension. OpenAI's Codex demonstrated outstanding performance in code generation and understanding, capable of handling multiple programming languages and generating high-quality code. In addition, Baidu and the Beijing Academy of Artificial Intelligence introduced Ernie 3.0 and CPM-2, further enriching the LMs ecosystem across different domains. Ernie 3.0 showed remarkable capabilities in multilingual processing, knowledge graph construction, and enhanced language understanding, whereas CPM-2 exhibited significant advantages in Chinese language generation and multimodal task processing. LMs came to the emergence stage from 2022 to 2024, with a large number of horizontal and vertical AI being developed and the diversification and specialization of LMs being witnessed. ChatGPT and InstructGPT worked on understanding and executing user commands, while GPT-NeoX and OPT innovated in model size and training efficiency. LMs, such as ChatGLM, BLOOM, and CodexGen2, continued to advance in multilingualism, programming language understanding, and code generation. Meanwhile, multimodal models such as the successor to DALL-E have demonstrated strong capabilities for the joint processing of images and texts.

Overall, LMs have demonstrated rapid changes and advanced in AI since 2018. This period witnessed model scale expansion, capabilities enhancement, and application scope broadening from preliminary LMs like GPT-1 to multimodal LMs such as GPT-4o in 2024. Diverse LMs such as InstructGPT, GPT-NeoX, and BLOOM embody technological breakthroughs within specific domains, while LMs for multimodal tasks mark a step toward more complex AI applications. The development of these LMs pushes the frontiers of natural language processing technologies and lays a solid foundation for future innovations. With the landing application of LMs, some problems are shown, including power consumption and catastrophic forgetting. Thus, solving these problems can help the further development of LMs and greatly empower their industrial applications. In addition, the continuous improvement of the multimodal processing capability of LMs should enable their applications to autonomous driving and other intelligent systems [7].

The rest of the paper is structured as follows. Section 2 discusses the classification of LMs. Section 3 describes the current problems of LMs, and Section 4 gives the solutions. Section 5 introduces LM's applications. Section 6 concludes this work.

2 | Classification of Large Models

Based on their input data, LMs can be categorized into three types: large language models, visual foundation models, and multimodal LMs.

2.1 | Large Language Models

Large language models are a class of deep learning models specialized for processing natural language [8]. They are capable of understanding, generating, and processing text by pretraining and fine-tuning. One example is the GPT family, such as GPT-4, which has demonstrated strong capabilities in natural language processing tasks. GPT-4 has achieved remarkable results in many applications, including text generation, and it can produce coherent and contextually accurate paragraphs, and perform machine translation, thereby facilitating seamless communication across different languages. Its advanced capabilities underscore the profound impact of large language models on natural language processing.

2.2 | Visual Foundation Models

Visual foundation models are specialized for processing images and videos [9]. They can recognize, classify, segment, and generate image content and perform well on visual tasks such as image classification and target detection [10]. For example, OpenAI has designed a text-to-video generation model named Sora. This model is trained extensively on vast data sets to generate realistic or imaginative videos based on textual commands provided by users. By leveraging advanced deep learning techniques, Sora can interpret complex textual descriptions and translate them into vivid and dynamic videos that capture the essence of the described scenarios. Its ability to simulate the physical world with precision and details demonstrates the incredible potential of

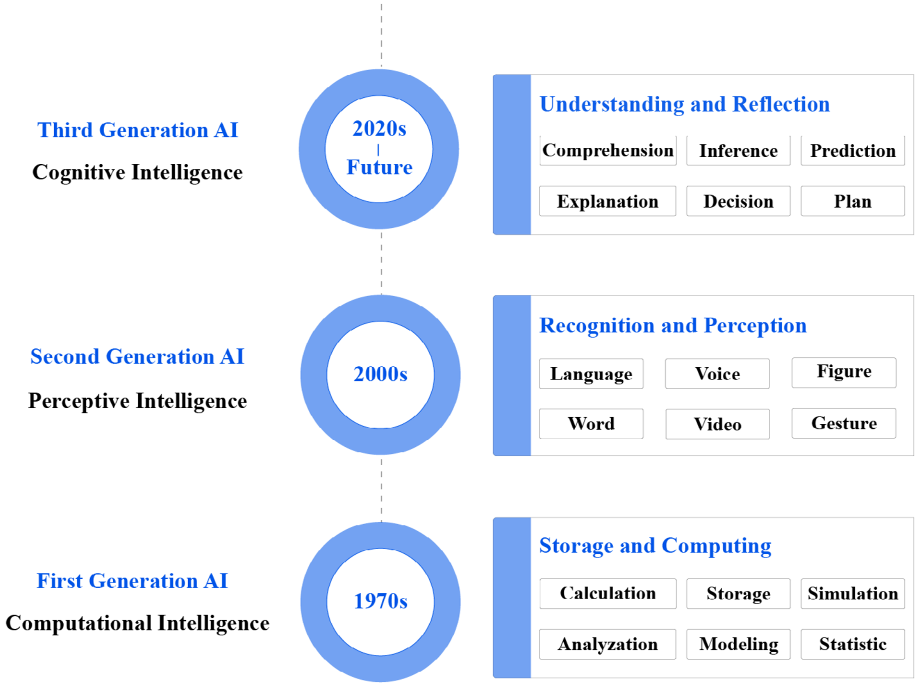

FIGURE 1 | Development of artificial intelligence.

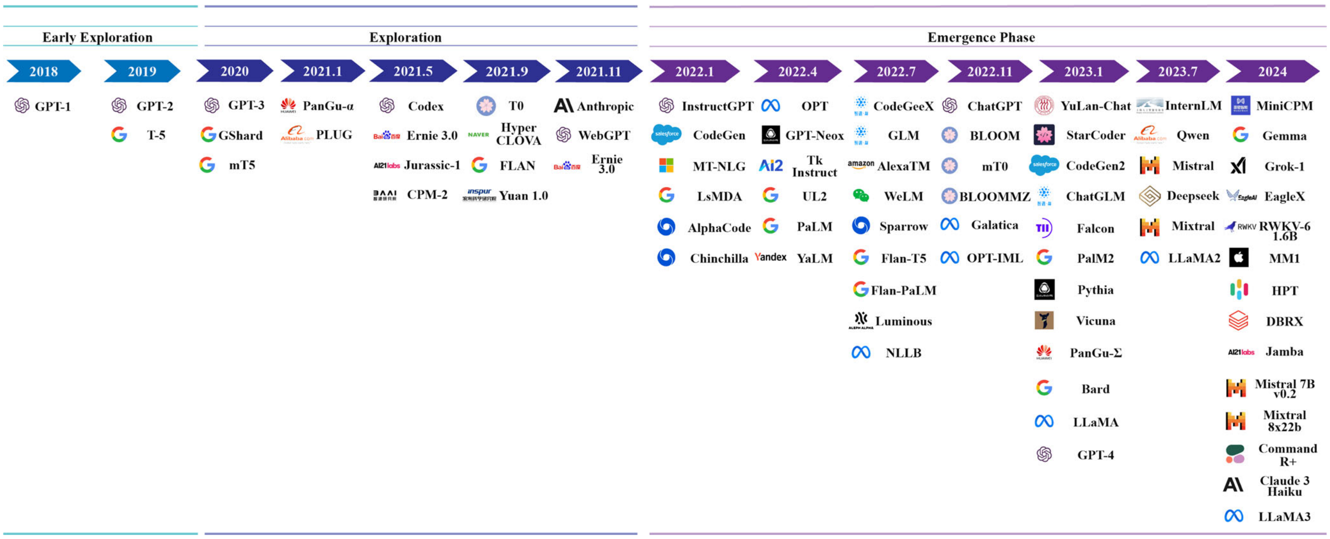

FIGURE 2 | Development of large models.

text-to-video generation models in various fields, including entertainment, education, and virtual reality.

2.3 | Multimodal Lms

Multimodal LMs can simultaneously process multiple data modalities [11], including text, images, and audio. These models combine linguistic and visual information to perform task processing with learned rich representations. One of the well-regarded multimodal language models is Contrastive Language-Image pretraining (CLIP), which has good performance across various visual and linguistic tasks. CLIP employs a comparative learning approach that involves pairing images with their corresponding textual descriptions and training the model to recognize and associate these pairs. By leveraging this methodology, CLIP has demonstrated remarkable proficiency in tasks such as image captioning, text-to-image retrieval, and visual question answering. Its ability to understand and relate visual and linguistic information makes it a valuable tool for advancing research in multimodal learning and fostering greater integration between visual and linguistic data.

These three types of LMs play vital roles in AI and are closely interconnected. Large language models and visual foundation ones are responsible for processing natural language and image data, and they can collaborate to solve complex multimodal tasks. Multimodal LMs combine linguistic and visual information and

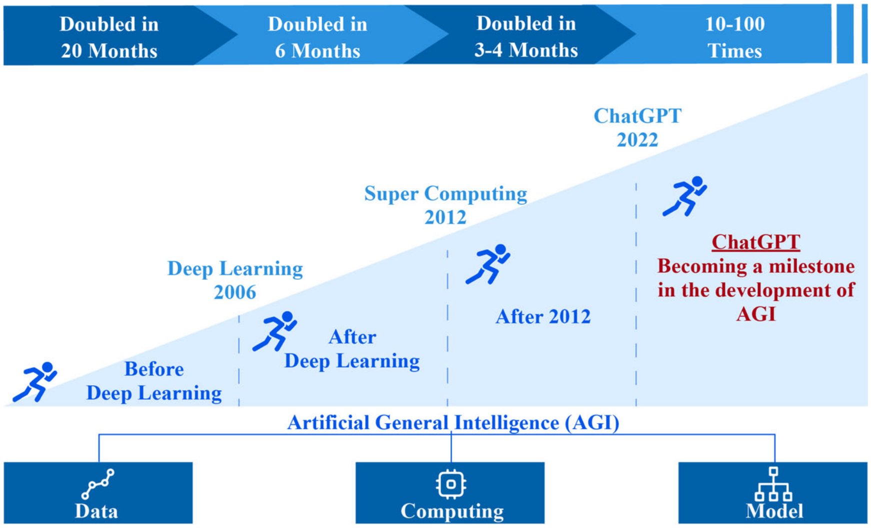

FIGURE 3 | Energy consumption of artificial intelligence in different phases.

can process multiple data modalities simultaneously, providing new possibilities for cross-modal intelligence tasks. In summary, these models are mutually reinforcing each other and complementary, driving the development and advancement of AI.

3 | Current Problems of Large Models

3.1 | Data and Energy Consumption

The more model parameters there are, the more data need to be fitted. LMs lead to excessive data consumption and computing power due to their large number of model parameters. The number of parameters of LMs has reached the trillion level. As noted in Figure 3, the energy consumption of AI is increasing. Hoffmann et al. [12] state that training data size and computational resource consumption are proportional to the number of parameters. This observation highlights a crucial challenge in the field of deep learning, where increasing the number of parameters to improve model performance often leads to a significant increase in both the amount of training data required and the computational resources needed for training. ChatGPT3 has 175 billion parameters, and its training data reach 300 billion words. More training data are required to fit models with more parameters. High-quality text data are expected to be exhausted by 2026 [13], posing a significant challenge for the continued advancement of large language models.

In the training process of LMs, the selection and quality of input data sources play a crucial role in the performance of the models. LMs rely on large-scale, diverse data sets, including different data types such as text, images, and audio. Primary data sources include data crawled from the Internet, publicly available databases, social media content, and proprietary data sets. However, the diversity and openness of these data sources also pose multiple challenges [14]. First, the accuracy and reliability of the data may vary, especially in web-crawled data, where the proportion of noise and misinformation is relatively high. Insufficient or underrepresented data for specific groups make the model inaccurate or even discriminatory. Second, the problem of data representativeness causes the model to perform poorly in dealing with particular tasks, especially when there is bias or underrepresentation of certain groups or topics in the training data. Therefore, when constructing an LM, data need to be carefully selected and cleaned to ensure that the data are diverse and of high quality, thereby leading to the LM's high generalization ability and reliability [15].

In addition, a deep learning model needs to consume far more energy than humans. AlphaGo played against a human master at Go in 2016. It needs to consume about 20,000 W of power, whereas a human Go player's brain power is only about 20 W. Furthermore, training ChatGPT3 on large cloud data centers requires 1,287,000 kWh of electricity [16], and it consumes 560,000 kWh of electricity daily during its problem-solving phase [17]. This high resource requirement hinders the development of LMs, triggering scholars to investigate more efficient methods, including parameter sharing, model pruning, and quantization. Some technical solutions for creating energy-efficient and sustainable data centers driven by AI methods are discussed in Buyya, Ilager, and Arroba [18]. These strategies aim to optimize the model structure for reducing model size and running costs.

3.2 | Catastrophic Forgetting

Training LMs on new tasks may impair their performance on previous tasks by failing to remember processed data or scenarios during the problem-solving phase [19]. This phenomenon is called catastrophic forgetting, and there is no effective mechanism to retain the model's learned knowledge and scenarios. In the training phase, fine-tuning the LM with domain data improves the model's performance in the vertical domain. However, it may impair performance in the general domain, especially when the new training data significantly differs from the original. The model may gradually lose its ability to accurately predict and interpret the original data over time, leading to a decline in performance. For instance, in unmanned driving, LMs for interpreting and responding to environmental data may fail to a) remember specific situations they have encountered and

b) implement more optimized decision-making controls when facing future similar scenarios. In such cases, the model's inability to effectively adapt or utilize experience can hinder its ability to cope with new and unforeseen situations, ultimately affecting the overall performance and reliability of unmanned driving systems [20].

3.3 | Reasoning Ability

The LM's logical reasoning ability is weak because it is a black-box model and cannot "divide and conquer." It performs poorly when dealing with complex problems that require logical and numerical reasoning. The inverse scale phenomenon [21] indicates that more parameters and training data lead to poor performance. These limitations indicate that increasing the model size does not solve all problems, especially for tasks requiring advanced cognitive and reasoning abilities. In addition, black-box reasoning processes are challenging to explain, and errors are difficult to correct. Researchers are exploring interpretable machine learning models that provide transparent decision processes and verifiable inference paths to address these challenges. This includes using decision trees, causal inference models, and open question-answering systems to enhance the logical reasoning of the model [22].

Despite the limited logical reasoning power of LMs, they can play an essential role in several critical aspects of the optimization problem. First, LMs can serve as a powerful tool for problem definition and formulation, helping one clarify the issue by translating natural language requirements into structured problem forms, thus providing precise inputs for subsequent optimization algorithms. In optimization strategy development, LMs can be an assistive design tool to help select appropriate algorithmic architectures and parameter settings by analyzing the literature and technical documents, providing rich information and best practices about optimization algorithms [23]. In addition, although LMs are usually not directly involved in the core computation, they can still be utilized in generating and screening candidate solutions to evaluate and interpret candidate solutions, thus providing useful heuristic guidance [24].

3.4 | Localization Fault

LMs often cannot recognize their errors or understand why they occur. They have difficulty in correcting them. This limitation makes it more challenging for LMs to fix them effectively [25]. For instance, in natural language processing tasks such as machine translation or sentiment analysis, LMs may produce incorrect outputs due to misunderstandings of the input text or biases in the training data. Without the ability to self-reflect and identify these issues, LMs cannot learn from their mistakes and evolve over time. This underscores the importance of developing more advanced error detection and correction mechanisms for LMs and ensuring that they are trained on diverse and unbiased data sets to minimize errors in the first place. This limitation restricts their performance in a wide range of real-world applications, as they cannot ensure the robustness of the results or point out the errors. Therefore, sophisticated feedback loops and error analysis techniques are developed [26]. These techniques are designed to help models accurately recognize and understand the nature of errors when feedback is received. Thus, it enables the model to make targeted adjustments. In addition, by combining human supervision with automatic learning mechanisms, the models can be enhanced in recognizing and correcting errors. This human–machine cooperative approach improves an LM's judgment and enhances its ability to self-adjust to deal with unknown challenges [27].

3.5 | Ethical Problems

The rapid development of AI has raised numerous ethical problems. Algorithmic bias is one of the most pressing ethical issues in this field [28]. LMs rely on patterns extracted from historical data during training, which often contain preexisting societal biases and inequalities. These biases can be reflected in the model's prediction results, influencing judgments about specific groups. For example, AI systems in recruitment, judicial decisions, and credit evaluations may result in discrimination against certain groups due to historical injustices in the training data. To minimize the effects of bias, developers need to review the data and take steps to ensure the fairness of AI systems [29].

Privacy protection issues are becoming more prominent with the popularity of AI and LM technologies [30]. LMs usually require a large amount of data for training. These data may contain sensitive information such as personal identity, behavior, and health status. This is especially prominent in the healthcare industry, where patient data are inevitably needed to train healthcare AI models. For example, in the case of facial palsy diagnosis, healthcare AI models require patient facial data for training [31]. Although patients sign a consent form when facial images are taken, ensuring patient data privacy in hospitals is still a problem. The public widely recognizes that AI technologies have the potential to revolutionize health care, leading to more accurate diagnoses [32]. However, social issues such as privacy and diagnostic accuracy need to be addressed, as people prefer human doctors to diagnose their health issues.

In LMs, if the data source is not transparent or the data are not processed properly, this may lead to the risk of user privacy breaches. In addition, user inputs may be stored or processed during the training or fine-tuning phase of LMs, and a lack of transparency about how these data are handled may raise privacy concerns because users are unsure whether their inputs are retained, how they are used, and whether a third party may access them. Furthermore, some LMs record the content of user interactions when providing their services, which could inadvertently collect sensitive information, leading to privacy violations or data breaches if these records are misused or stored unencryptedly. Inadequate authorization or improper handling of private data may lead to privacy leakage and violation of individuals' privacy rights. This problem has triggered ethical controversies about data collection, storage, and use [33]. In addition, security flaws in data storage and transmission may make these data targets for cyberattacks. Most people are concerned about the misuse of personal data and believe that users' consent and right to information should be prioritized in data use. To address these privacy challenges, it is essential to implement data anonymization, encrypt transmission, and provide transparent privacy policies to enhance public trust and protect user privacy.

Finally, the issue of liability is another important ethical consideration. As AI plays an increasingly important role in decision-making, clarifying where responsibility lies becomes particularly complex. For example, when a mechanical assembly line system adjusts its processes and makes wrong decisions due to AI, should the AI developer, user, or system itself be held responsible? A legal framework is needed to clarify the responsibilities and obligations of all parties in such technologies. Thus, they can be kept accountable when problems arise [34]. Another example is that if an AI model is applied to medical decision-making, it may cause serious consequences if it misdiagnoses a patient's health issues. Therefore, in the case of applications that require accurate results, the judgment and output of LMs need to be judged and processed by manual models.

3.6 | Fundamental Problem

3.6.1 | Network Architecture Flaws

The combination of different modules and the hierarchical structure of the LMs architecture lacks clear functional definitions and connections to human-understandable knowledge of mechanisms [35]. This architecture cannot learn the causal relationships for reasoning, limiting an LM's performance in performing complex tasks. In addition, it cannot effectively handle dynamic changes and new situations due to poor plasticity and interoperability. Current LMs are based on the Transformer architecture as the underlying model, which can handle the simultaneous text input from a large number of users. However, if a large number of users simultaneously input audio or even video as tokens into a multimodal LM, it may be difficult for the LM to respond to the users promptly. This causes queuing and waiting or even crashes of the model, which affects the user experience [36]. Therefore, optimizing the processing capability and architectural design of an LM is necessary to better cope with the challenge of multimodal input. In addition, the stability and responsiveness of an LM under high load conditions should be improved.

3.6.2 | Training and Reasoning Defect

Autoregressive and autoencoder are commonly used to train LMs. The core is a back-propagation mechanism that takes the global error as the optimization goal and iteratively updates all parameters. This process consumes a large amount of training data and computing power to optimize an LM's performance on a specific task. It is prone to overfitting and reduced model flexibility [37]. When facing new tasks, LMs often need to be adapted and learned from the beginning, which is inefficient and imposes limitations on the long-term development of LMs. Therefore, training yields all parameters that correspond to a specific task. When learning a new task, the backpropagation mechanism updates all parameters, leading to forgetting the old task. This catastrophic forgetting highlights the limitations of existing models in continuous learning and adapting to new environments. There is an urgent need to enhance the generalization ability and adaptability of LMs by improving training and optimization strategies.

In the reasoning phase, an LM is guided by the question and parameters obtained from training. The forward reasoning of answering a question requires all layers and all parameters to be involved in the computation, consuming a large number of computing resources. It is estimated that the energy overhead of a single ChatGPT interaction is more than a hundred times that of a Google search. This implies that energy consumption becomes a major issue in large-scale deployment and frequent use scenarios [38]. Exploring more efficient model structures and algorithms to reduce energy consumption [39] and maintain performance becomes especially critical.

4 | Potential Solutions

4.1 | Increasing Data and Computation Capability

Sutton [40] point out that all the skills of AI cannot compare to the powerful computing power and generalized algorithms, and the key to the progress of AI lies in the computing power and data. Kaplan et al. [41] indicate that scale is important. More data and bigger models bring better results. Emerging properties [42] show that as models grow in size and data volume, LMs can suddenly emerge with capabilities that were not previously available. This "emergent" capability suggests that complex behaviors and functions can be observed, which are not visible or attainable in smaller models. To take full advantage of these emergent properties, researchers have explored how to design and train large-scale models more efficiently by using, for example, more advanced optimization algorithms [43] and finer-grained parameter tuning strategies. These efforts aim to advance AI technologies and ensure that these advanced AI systems work safely and reliably in real-world applications.

In addition, focusing only on scaling up models is unrealistic. To better support the concurrent access of a large number of users, a centralized data center is difficult to support. Therefore, reducing the model size while decentralizing computing to the edges and terminals is also important. Through model optimization, centrality, and distributed computing, LMs can be executed on edge devices, including smartphones, devices of the Internet of Things, and automotive computers. This meets user needs more efficiently, enhances data privacy, and reduces latency by bringing computation closer to data sources. Edge devices are often limited by computational capability, memory, storage, and energy consumption, and directly deploying LMs poses challenges [44]. In that case, these limitations require improving model performance and reducing computational resource consumption. One possible approach includes model compression techniques such as quantization, pruning, and knowledge distillation. Quantization reduces a model's storage and computation requirements by converting floating-point numbers to low-precision representations. Pruning simplifies the model structure by removing unimportant weights [45]. Knowledge distillation utilizes the knowledge generated by an LM to train a smaller model, significantly reducing the model's size while maintaining its performance [46]. In addition, cloud-assisted mobile edge computing is an effective strategy to facilitate model running on edge devices. Edge devices can complete preprocessing, feature extraction, and other preliminary work and offload complex computational tasks to the cloud, thus reducing the burden on edge devices.

4.2 | Neural-Symbolic Synergy

The advantages of neural networks are their abundant prior knowledge, strong generalization, and excellent flexibility [47], but the disadvantages are their weak inference, poor interpretability, and presence of illusions [48]. Symbolic rules have the advantage of combinability, interpretability, and strong higher-order reasoning [49], but the disadvantage of combinatorial explosion, noise sensitivity, and poor generalization [50]. To overcome these limitations, existing research attempts to combine the strengths of these two systems to explore a new hybrid paradigm in a complementary manner. There are five possible combination paradigms: (A) symbolic systems as the architecture and neural networks as a subroutine; (B) neural networks that convert nonsymbolic inputs into symbols, which symbolic systems can process; (C) neural networks trained with symbolic rule data to realize symbolic rule functions; (D) structural templates in neural networks based on symbolic rules; and (E) iterative interaction between neural networks and symbolic reasoning. With the above strategies, LMs can accurately simulate human cognitive processes and provide higher operational flexibility and decision quality, especially in application scenarios that require complex reasoning.

4.3 | Data Orientation to Human Pattern

The development of neural networks draws on brain science, creating the foundation of the current LMs. If the mechanisms of memory representation, activation, retrieval, encoding, and playback in brain science can be used in LM's development, it is expected to break the inherent defects of the current LMs. Figure 4 shows that brain science corresponds to the development of AI. It is shown that each breakthrough in cognitive brain science corresponds to a breakthrough in AI.

Memory is the foundation of human intelligence. It influences the human brain's intellectual activities, including learning, abstraction, association, and reasoning. These activities are influenced by encoding, storage, and retrieval. Encoding organizes and transforms external information. Learning efficiency depends on memory encoding strategies [51]. Therefore, multi-channel encoding and situational association strategies can significantly improve learning effectiveness. Storage saves learned knowledge in long-term memory in hierarchical categories, retaining gained knowledge such that subsequent learning can be more efficiently aided by prior experience. Retrieval extracts information from long-term memory, consolidates memory storage, stimulates metacognitive abilities, and promotes reasoning, abstraction, and association. The human memory model [52] is shown in Figure 5. Memory retrieval and activation are the basics of intelligence. Unlike LMs, where all parameters need to be involved in reasoning, the brain retrieves a small amount of knowledge from long-term memory and turns it into working memory for reasoning through an activation mechanism. The human brain contains about 10^{11} neurons and 10^{15} synapses. They only consume 20–23 W of energy, whereas an LM of the same size consumes up to 7.9×10^6 W of energy [53]. Thus, memory's retrieval and activation mechanism brings inspiration for designing machine intelligence models and representation mechanisms, which is expected to break through the shortcomings of LMs that overconsume data and computing resources.

5 | Large Model'S Applications

In the technology ecosystem, LMs need to be equipped with powerful natural language processing capabilities and need to be integrated into a variety of applications for interoperability seamlessly [54]. This interoperability allows LMs to be widely used on different platforms, thus better serving diverse user needs and enhancing their application value. Through interoperability with common applications, LMs can be embedded into productivity tools, customer service systems, educational platforms, and other scenarios to provide intelligent interaction and decision support. Defining standardized interface protocols [55] can promote interoperability among LMs and general purpose applications. In that case, LMs can be seamlessly invoked as service modules by various applications. For example, LMs can use web service technologies to handle data requests from different applications and realize cross-platform and cross-language data exchange and processing. In addition, the deep integration of LMs with applications is also essential. By embedding the LM software development kit, applications can realize tighter data flow control and real-time response. This approach reduces the latency of data transmission. It enables the model to be customized and optimized according to the specific needs of an application, further enhancing user experience and system performance. The realization of this interoperability greatly expands the application capability of LMs, providing a more intelligent, efficient, and personalized service experience.

However, one critical aspect of developing vertical domain LMs on top of some base LMs is ensuring the data relevance and quality for fine-tuning, as this directly impacts their performance and accuracy. Ensuring that an LM maintains its generality while adapting to tasks in specific domains poses a challenge during the fine-tuning process. As the scale of the model increases, its complexity and lack of interpretability also increases. This may lead to difficulties in understanding and controlling the model's outputs in practical applications. Furthermore, LMs for vertical domains need to be integrated with existing enterprise processes and possess collaboration and controllability capabilities. This requires LMs to be embeddable within existing systems, assisting in upgrading specific components rather than serving as a full replacement. Therefore, their ability to interface with existing business personnel or systems needs to be focused when practical applications are handled. LMs for vertical domains require attention to knowledge base maintenance and updates. Businesses frequently change and maintain their knowledge bases, and the materials within these knowledge bases are diverse. Ensuring that LMs can efficiently and accurately identify relevant knowledge within different knowledge base systems and subsequently retrieve high-quality answers is a challenge that needs to be addressed. This section introduces three typical real-world applications of LMs, including autonomous driving, smart industrial productions, and intelligent medical assistance.

5.1 | Autonomous Driving Technologies

Autonomous driving technologies, such as automatic parking, adaptive cruise control, and traffic congestion assistance systems, can be used to facilitate people's daily travel. These technologies improve the safety and comfort of driving and effectively reduce

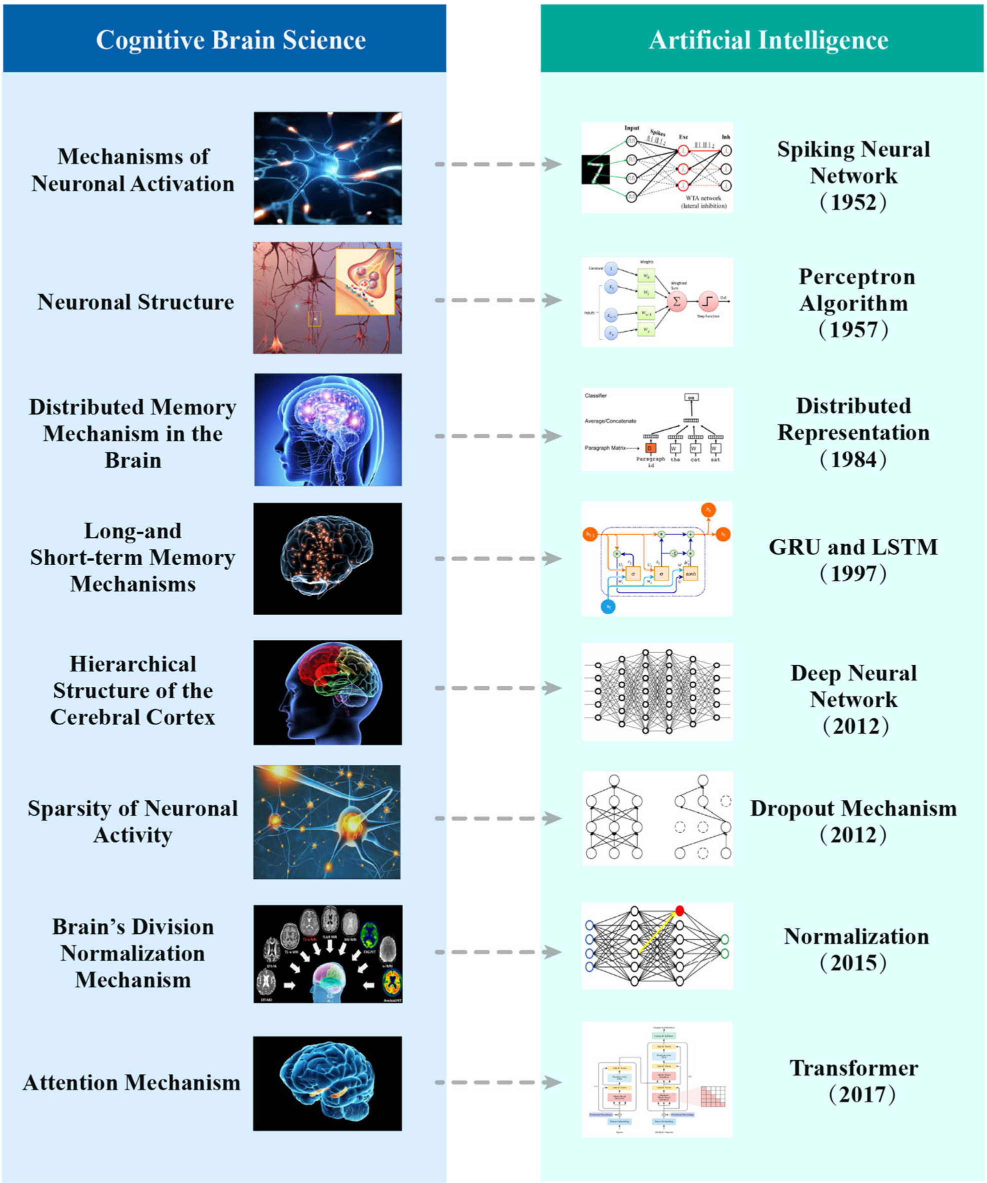
Cognitive Brain Science
Artificial Intelligence
Mechanisms of Neuronal Activation
Spiking Neural Network (1952)
Neuronal Structure
Perceptron Algorithm (1957)
Distributed Memory Mechanism in the Brain
Distributed Representation (1984)
Long-and Short-term Memory Mechanisms
GRU and LSTM (1997)
Hierarchical Structure of the Cerebral Cortex
Deep Neural Network (2012)
Sparsity of Neuronal Activity
Dropout Mechanism (2012)
Brain's Division Normalization Mechanism
Normalization (2015)
Attention Mechanism
Transformer (2017)

FIGURE 4 | Cognitive brain science versus artificial intelligence.

traffic accidents. Automated driving has experienced the following five steps of development [56].

1. Momentary Driver Assistance: The driver is fully responsible for driving the vehicle while the system provides momentary assistance like warnings and alerts or emergency safety interventions.
2. Driver Assistance: The driver is fully responsible for driving the vehicle while the system provides continuous assistance with either acceleration/braking or steering.
3. Conditional Automation: The system handles all aspects of driving while the driver remains available to take over driving if the system can no longer operate.
4. High Automation: When engaged, the system is fully responsible for driving tasks within limited service areas. A human driver is not needed to operate the vehicle.
5. Full Automation: When engaged, the system is fully responsible for driving tasks under all conditions and roadways. A human driver is not needed.

Humans use a dual-process system when driving, which is the basis for humans to realize complex reasoning [57]. Dual-process theory suggests that the human brain has dual-system synergistic mechanisms and multiple memory types that can support complex reasoning, including logic/intuition, implication/association, and explicit/implicit. It is shown in Figure 6 that the first intuitive thinking system is fast and does not consume computational resources, for example, accelerating and applying brakes. The second rational thought system is required to analyze and consume computational resources, for example, path planning. Current autonomous driving technologies cannot simulate the first system, resulting in low efficiency. To make LMs' decisions that can more closely resemble a human dual-process

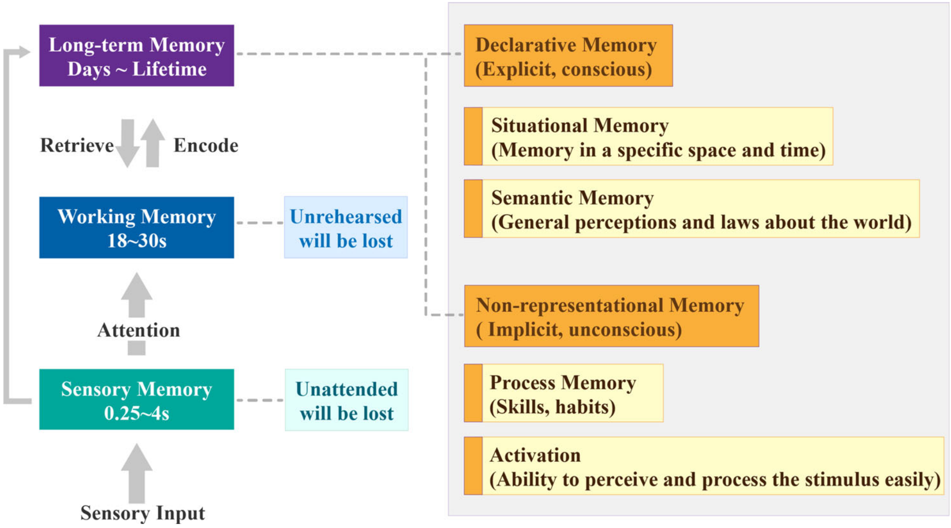

FIGURE 5 | Human memory model.

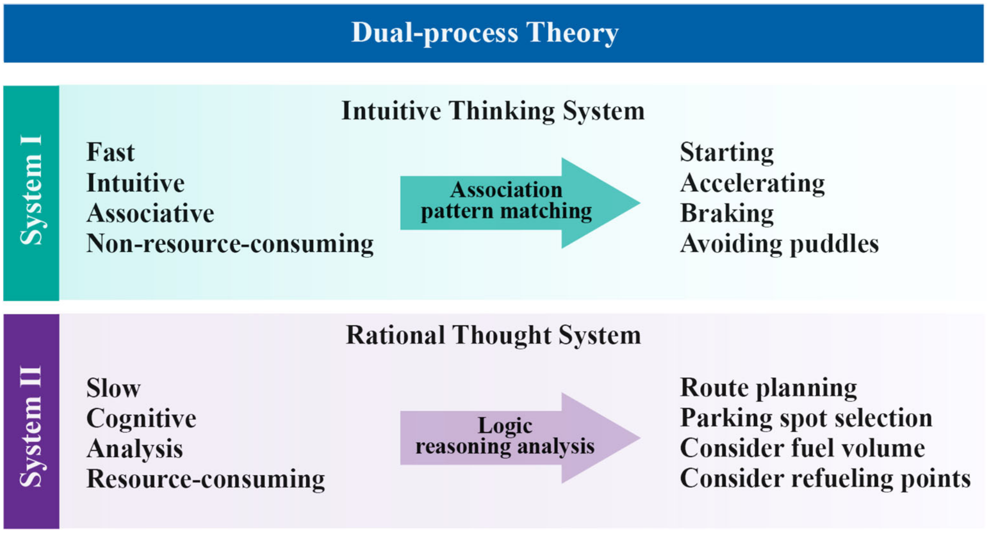

FIGURE 6 | Dual process theory in driving.

system, Tesla [58] proposes the Full-self Driving (FSD) system as a new “end-to-end autonomous driving.” It uses a neural network exclusively for vehicle control, which controls everything from computer vision to driving decisions. This neural network is trained from millions of video clips, replacing over 300,000 lines of C++ code. Thus, it reduces the vehicle system’s reliance on code and brings it closer to the human driver’s decision-making process. However, the LM still struggles to remember specific road conditions encountered in the past and cannot implement a more optimized decision-making strategy when similar scenarios are encountered again. Future LMs on simulating human memory systems are hopeful to solve this problem.

Driven by LMs and AI, automated driving technologies and the smart car industry are gradually breaking through the bottleneck of traditional technologies and ushering in a new phase of rapid development. Using advanced AI technologies, autonomous driving systems can better cope with complex traffic environments and improve driving safety and efficiency. With the continuous investment of major technology companies and automakers, autonomous driving technologies based on LMs is expected to realize wider applications and revolutionize how we travel. Its advancement is a revolution in the automotive industry and should profoundly impact the future urban transportation system and people’s daily lives.

5.2 | Smart Industrial Productions

LMs are widely applied in smart industrial productions [59]. They support industrial text generation and industrial knowledge Q&A to realize the generation of production handover reports, equipment point inspection records, etc. Industrial multimodal LMs accomplish helmet detection, statistics of product defective rate, etc. They greatly improve the efficiency of workers’ traditional manual processing. The framework of industrial LMs is shown in

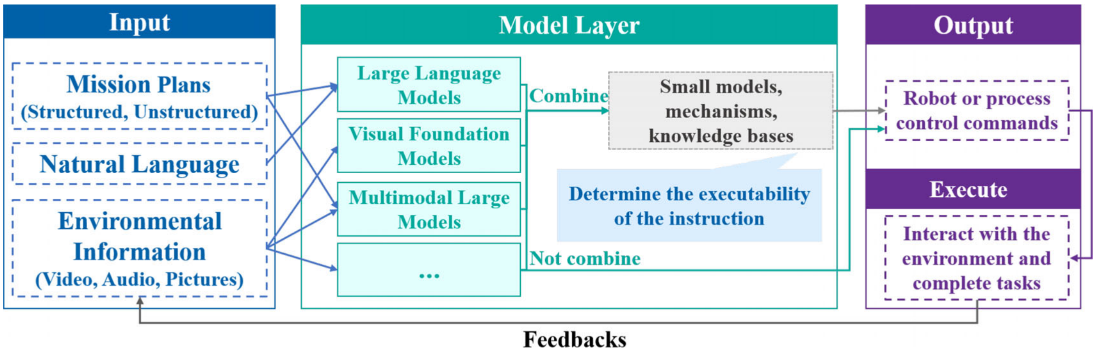
Input
Mission Plans
(Structured, Unstructured)
Natural Language
Environmental Information
(Video, Audio, Pictures)
Model Layer
Large Language Models
Visual Foundation Models
Multimodal Large Models
...
Combine
Small models, mechanisms, knowledge bases
Determine the executability of the instruction
Not combine
Output
Robot or process control commands
Execute
Interact with the environment and complete tasks
Feedbacks

FIGURE 7 | Structure of industrial large models.

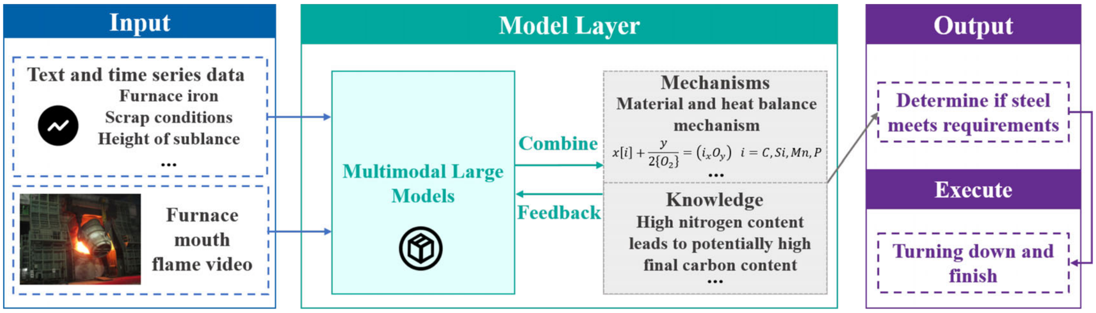
Input
Text and time series data
Furnace iron
Scrap conditions
Height of sublance
...
Furnace mouth flame video
Model Layer
Multimodal Large Models
Combine
Feedback
Mechanisms
Material and heat balance mechanism
$x[i] + \frac{y}{2\{O_2\}} = (i_x O_y) \quad i = C, Si, Mn, P$
...
Knowledge
High nitrogen content leads to potentially high final carbon content
...
Output
Determine if steel meets requirements
Execute
Turning down and finish

FIGURE 8 | Structure of the converter steelmaking large model.

Figure 7. Specifically, the input includes structured and unstructured mission plans, natural language, and multimodal environmental information. They are input to the model layer, and an LM is selected according to the task type. After LM processing, small models, mechanisms, and knowledge bases are possibly employed to make judgments about the correctness of results, preventing erroneous instructions generated by LM from being executed. After executing the instructions, the new state is fed into the LM again as environmental information. This framework is applicable to a large number of industrial scenarios.

Take the converter steelmaking scenario as an example. Accurate prediction of the endpoint steel's carbon level and temperature is essential to ensure the steel-making success rate. Traditional methods use physicochemical change modeling to predict endpoint carbon content and temperature through a mechanistic approach. However, the accuracy of traditional prediction methods is difficult to guarantee due to the uncertainty of the reaction in the furnace and the composition of the added pig iron. Multimodal LMs provide a solution because of their robust multimodal processing capabilities. They combine flame characteristics of the furnace with sensor data to make real-time judgments about the carbon content and temperature in the furnace. Figure 8 shows the structure of the converter steelmaking LM. It utilizes multimodal data as inputs, including textual data (furnace iron, scrap conditions, etc.), time-series data (height of sublance, flow rate of oxygen blowing, etc.), and furnace mouth flame video. Then, the above multimodal data are processed, and the LM predicts endpoint carbon content and temperature. However, the initial training of the LM results in an inaccurate model due to the small amount of training data. Thus, mechanisms and knowledge related to converter steelmaking are combined to assist the LM. Finally, as data continue to accumulate, the predictive accuracy of the LM is further improved.

In addition to the steel industry, LMs have a wide range of applications in other industrial fields [60], such as textile, metallurgy, and food industries. They can assist in optimizing a design process, expand manufacturing intelligence, improve operation and management, and promote the intelligence of products and services.

5.3 | Intelligent Medical Assistance

The application of LMs in the medical field is becoming increasingly prevalent. It is shown in Figure 9 that some medical procedures are becoming more intelligent by introducing LM technologies into vertical healthcare domains. The construction of clinical intelligence is better supported by the fusion of multimodal data, including text, images, and time-series data. It is shown in Figure 9 that applications include disease diagnosis and decision support, disease prediction and prevention, medical imaging analysis, pharmaceutical research and development, medical literature analysis and knowledge discovery, intelligent medical Q&A and consultation, healthcare process optimization, medical education and training, and hospital management. Some applications are well applied and serve the public, such as intelligent medical Q&A and hospital management. The former can answer people's questions about common diseases and guide treatment [61]. The latter can significantly reduce the queuing time in hospitals and increase the public's satisfaction with AI applications in health care.

1. Disease Diagnosis and Decision Support: In clinical decision support, LMs provide intelligent diagnosis and

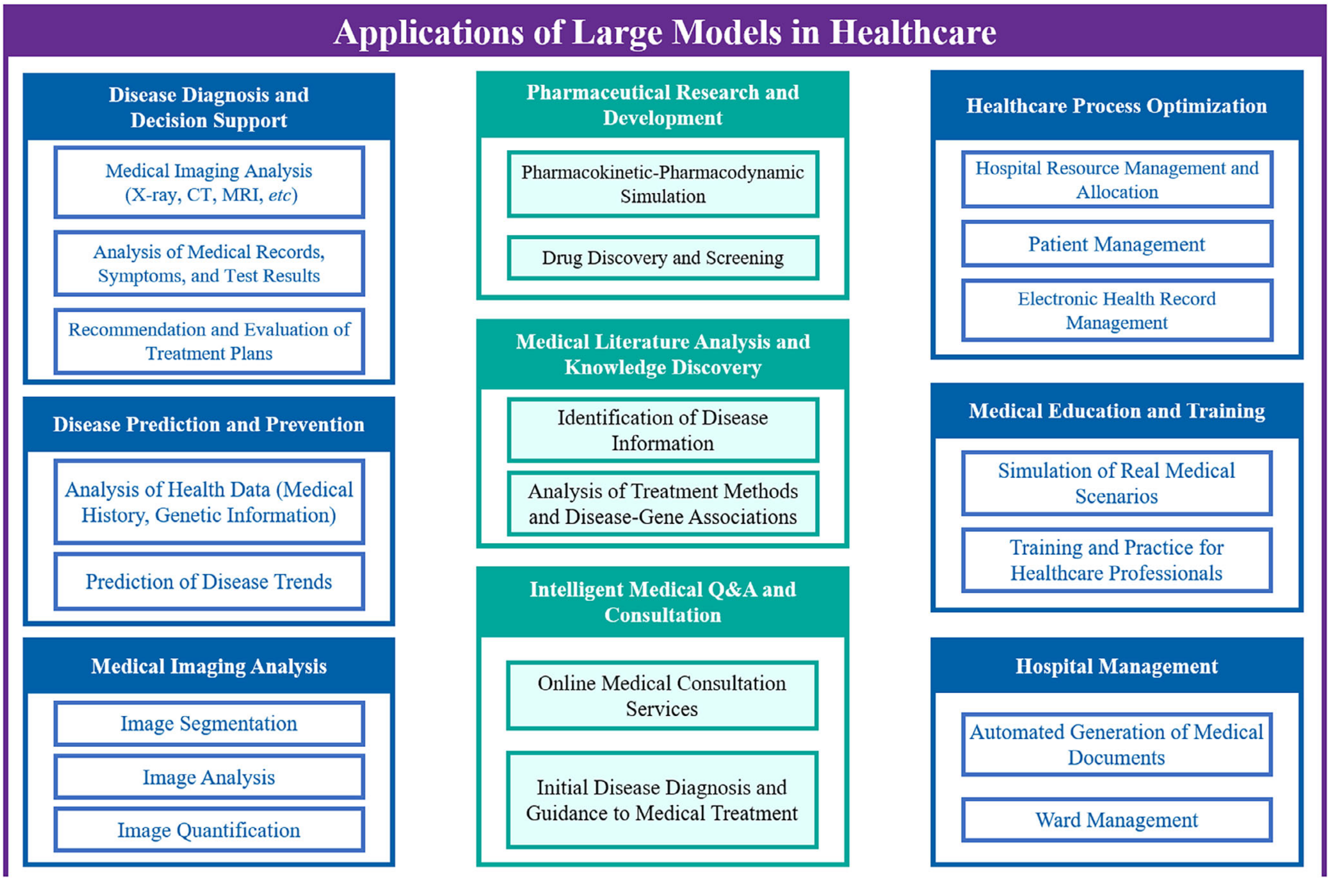
Applications of Large Models in Healthcare
Disease Diagnosis and Decision Support
Medical Imaging Analysis (X-ray, CT, MRI, etc)
Analysis of Medical Records, Symptoms, and Test Results
Recommendation and Evaluation of Treatment Plans
Disease Prediction and Prevention
Analysis of Health Data (Medical History, Genetic Information)
Prediction of Disease Trends
Medical Imaging Analysis
Image Segmentation
Image Analysis
Image Quantification
Pharmaceutical Research and Development
Pharmacokinetic-Pharmacodynamic Simulation
Drug Discovery and Screening
Medical Literature Analysis and Knowledge Discovery
Identification of Disease Information
Analysis of Treatment Methods and Disease-Gene Associations
Intelligent Medical Q&A and Consultation
Online Medical Consultation Services
Initial Disease Diagnosis and Guidance to Medical Treatment
Healthcare Process Optimization
Hospital Resource Management and Allocation
Patient Management
Electronic Health Record Management
Medical Education and Training
Simulation of Real Medical Scenarios
Training and Practice for Healthcare Professionals
Hospital Management
Automated Generation of Medical Documents
Ward Management

FIGURE 9 | Applications of LMs in healthcare.

treatment recommendations for physicians by integrating and analyzing vast amounts of medical literature, guidelines, and case data. When confronted with complex medical conditions, LMs can rapidly retrieve relevant clinical trials or case studies, assisting physicians in evaluating potential effect and risk of different treatment options, thereby enabling more precise clinical decision-making [62]. Furthermore, multimodal AI-assisted diagnosis can assess patients' disease status from multiple dimensions and provide physicians with more diagnostic information and reference points, which is crucial for clinical decision-making.

2. Disease Prediction and Prevention: By leveraging the powerful data processing capabilities and pattern recognition techniques of LMs, we can conduct the in-depth analysis of vast epidemiological data and medical records, enabling precise predictions of disease transmission trends and risk assessments [63]. For instance, through training on extensive case data, LMs can identify subtle abnormalities in X-rays or CT scans, assisting physicians in making more accurate diagnoses. Furthermore, they can assess an individual's risk of viral infection based on their health data, lifestyle habits, medical history, and other information, providing a scientific basis for formulating personalized epidemic prevention measures. Such personalized prevention strategies help reduce the risk of epidemic transmission, enhance the utilization efficiency of medical resources, and alleviate the burden on the healthcare system.

3. Medical Imaging Analysis: LMs are progressively becoming the backbone of Medical Imaging Analysis. Taking tumor detection as an example, through training on vast amount of imaging data, including CT and MRI, LMs can automatically identify and analyze essential information, such as tumor morphology, size, and location in images, with an accuracy that surpasses traditional manual image review methods. For instance, deep learning models developed by the Google's DeepMind team has achieved significant results in analyzing ocular OCT images, aiding doctors in early screening and diagnosis of ophthalmic diseases [64]. In addition, LMs can adapt to multimodal medical data, including X-rays, CTs, MRIs, PETs, ultrasounds, pathology slides, endoscopes, dermatoscopes, dental radiographs, blood smears, and more, enabling identification at levels ranging from organs to lesions, and even down to pathological cells, diseased tissues, and cellular elements. In the early detection of lung cancer, LMs can accurately identify minute lung nodules by analyzing low-dose CT scan images, gaining precious treatment time for patients. Furthermore, they can process static images and analyze time series of imaging data, such as dynamic MRIs, providing doctors with more comprehensive disease information.

4. Pharmaceutical Research and Development: LMs play a significant role in new drug discovery and development.

In drug molecule design, the successful application of AlphaFold in protein structure prediction [65] provides robust support for structure-based drug design. LMs have the potential to accurately predict three-dimensional structures of proteins, assisting researchers in understanding interaction mechanisms between drugs and targets, thereby enabling the design of drug molecules with excellent selectivity and activity. By simulating the interactions between drugs and biomolecules, LMs can predict drugs' potential effects and side effects, thereby accelerating the drug screening process and reducing research and development cost. For example, with the assistance of Huawei Cloud's Pangu LM [66], the screening process for drug molecules has been dramatically accelerated, shortening the lead drug research and development cycle from several years to approximately 1 month. The research and development cost has been reduced by about 70%, and the super antibiotic cinnamaldehyde has been successfully developed.

5. Medical Literature Analysis and Knowledge Discovery: By leveraging the powerful natural language processing and deep learning capabilities of LMs, the ways medical knowledge is acquired, integrated, and applied are transformed. These models can efficiently process vast amounts of medical literature data, extract essential information, and uncover potential knowledge patterns, providing robust support for medical research. BioBERT [67] is a bidirectional encoder model based on the Transformer architecture, specifically designed for biomedical text mining. By training on a large corpus of biomedical literature, it can understand and process specific vocabulary and concepts in the biomedical field, making precise analysis of medical literature possible. In the medical knowledge discovery, LMs have demonstrated exceptional performance in medical testing. They can understand medical questions and guide users in deep thinking through dialogue, helping them discover new medical knowledge and insights. By constructing a medical corpus that covers multidisciplinary medical literature, textbooks, guidelines, reports, patient records, and other data, LMs can excel in structured and standardized medical knowledge question-answering and various clinical text analysis tasks such as report quality control. They can also undertake forward-looking explorations such as diagnostic predictions.

6. Intelligent Medical Q&A and Consultation: Chatbots powered by LMs can also play a significant role in medical consultations and health inquiries. These intelligent assistants can answer patients' common questions about their conditions, medications, and side effects, enhancing patient experience. Furthermore, when necessary, they can direct patients to appropriate medical facilities or specialists for further consultation, optimizing the allocation of medical resources.

7. Healthcare Process Optimization: In electronic health record management, LMs can automatically extract and organize critical information from patients' medical histories, thereby reducing the paperwork burden on healthcare professionals and enhancing the readability and retrieval efficiency of the data. By leveraging natural language processing techniques, LMs can quickly and accurately identify patients' medical histories, allergy histories, and current medication statuses, providing doctors with comprehensive yet concise patient profiles [68].

8. Medical Education and Training: LMs can simulate realistic medical scenarios and cases, providing medical students with immersive learning experiences. They can also guide students to think and explore actively through dialogue and interaction, cultivating their clinical thinking and decision-making abilities. Furthermore, these models can intelligently generate case studies and examination questions to direct students to complete the analysis of related subjects. Adopting a full-scenario and full-process application path that integrates clinical practice, data collection, scientific research, and teaching creates an innovative model for clinical diagnosis, data acquisition, storage and management, database construction, scientific research model design, and scientific research service support.

9. Hospital management: In medical documentation generation, LMs have demonstrated the ability to automatically produce high-quality admission/discharge summaries, case reports, surgical records, and other medical documents. This not only enhances the accuracy and standardization of medical documentation but also significantly reduces the workload of healthcare professionals. Furthermore, these models can generate personalized treatment plans and nursing protocols based on patients' specific conditions and doctors' instructions, providing patients with more precise and efficient medical services. In ward management, LMs can achieve intelligent bed allocation, patient tracking, and nursing resource scheduling by integrating and analyzing multidimensional data, including patients' medical histories, examination and test results, and medication records. This improves the efficiency of ward utilization and patient satisfaction. Additionally, LMs can predict patients' medical needs, providing timely alerts and decision support for healthcare professionals, enhancing the hospital's management level and medical service quality. With technological advancements, LMs are poised to achieve more innovations in the field of medical assistance, continuously enhancing the quality and efficiency of global medical services and bringing more personalized and precise healthcare services to society.

6 | Conclusions

The rapid development of LMs is reshaping many industries because they can better mine the value of data to accomplish assisted decisions. However, LMs face a number of unsolved problems, including data and energy consumption, catastrophic forgetting, and poor reasoning ability, which limit their applications. This paper has introduced the development of LMs. Then, the current problems of LMs are identified, and potential solutions are discussed. Finally, the applications of LMs in autonomous driving, smart industrial productions, and intelligent medical assistance are discussed. We believe that LMs have a promising future and can empower the development of various industries.

Author Contributions

Jing Bi: conceptualization, methodology, writing- original draft preparation, project administration. **Ziqi Wang:** visualization, investigation. **Haitao Yuan:** supervision, data curation, investigation, funding acquisition. **Xiankun Shi:** validation. **Ziyue Wang:** validation. **Jia Zhang:** formal analysis, writing- reviewing and editing. **MengChu Zhou:** writing-reviewing and editing. **Rajkumar Buyya:** writing- reviewing and editing.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data sharing is not applicable to this article as no data sets were generated or analyzed during the current study.

References

1. H. Lin and C. Tang, "Analysis and Optimization of Urban Public Transport Lines Based on Multiobjective Adaptive Particle Swarm Optimization," *IEEE Transactions on Intelligent Transportation Systems* 23, no. 9 (2022): 16786–16798.

2. X. Luo, H. Qu, Y. Wang, Z. Yi, J. Zhang, and M. Zhang, "Supervised Learning in Multilayer Spiking Neural Networks With Spike Temporal Error Backpropagation," *IEEE Transactions on Neural Networks and Learning Systems* 34, no. 12 (2023): 10141–10153.

3. G. Calafiore, S. Gaubert, and C. Possieri, "A Universal Approximation Result for Difference of Log-Sum-Exp Neural Networks," *IEEE Transactions on Neural Networks and Learning Systems* 31, no. 12 (2020): 5603–5612.

4. Q. Zheng, Z. Peng, Z. Dang, et al., "Deep Tabular Data Modeling With Dual-Route Structure-Adaptive Graph Networks," *IEEE Transactions on Knowledge and Data Engineering* 35, no. 9 (2023): 9715–9727.

5. J. Ruan, Q. Zheng, R. Zhao, and B. Dong, "Biased Complementary-Label Learning Without True Labels," *IEEE Transactions on Neural Networks and Learning Systems* 35, no. 2 (2024): 2616–2627.

6. L. Zhang, S. Wang, J. Liu, et al., "MuL-GRN: Multi-Level Graph Relation Network for Few-Shot Node Classification," *IEEE Transactions on Knowledge and Data Engineering* 35, no. 6 (2023): 6085–6098.

7. H. Muslim, "Design and Evaluation of Lane-Change Collision Avoidance Systems in Semi-Automated Driving," *IEEE Transactions on Vehicular Technology* 72, no. 6 (2023): 7082–7094.

8. X. Kong and Z. Ge, "Deep PLS: A Lightweight Deep Learning Model for Interpretable and Efficient Data Analytics," *IEEE Transactions on Neural Networks and Learning Systems* 34, no. 11 (2023): 8923–8937.

9. Z. Peng, M. Luo, W. HUang, et al., "Learning Representations by Graphical Mutual Information Estimation and Maximization," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, no. 1 (2023): 722–737.

10. M. Viggiato and C. Bezemer, "Leveraging the OPT Large Language Model for Sentiment Analysis of Game Reviews," *IEEE Transactions on Games* 16, no. 2 (2024): 493–496.

11. H. Han, Q. Zheng, M. Luo, K. Miao, F. Tian, and Y. Chen, "Noise-Tolerant Learning for Audio-Visual Action Recognition," *IEEE Transactions on Multimedia* 26, no. 2 (2024): 7761–7774.

12. J. Hoffmann, S. Borgeaud, A. Mensch, et al., "Training Compute-Optimal Large Language Models," *36th International Conference on Neural Information Processing Systems* (2022): 30016–30030.

13. F. Xue, Y. Fu, W. Zhou, Z. Zheng, and Y. You, "To Repeat or Not to Repeat: Insights From Scaling LLM Under Token-Crisis," *36th International Conference on Neural Information Processing Systems* (2022): 59304–59322.

14. R. Hai, C. Koutras, C. Quix, and M. Jarke, "Data Lakes: A Survey of Functions and Systems," *IEEE Transactions on Knowledge and Data Engineering* 35, no. 12 (2023): 12571–12590.

15. Q. Li, Z. Li, Z. Zheng, et al., "Capitalize Your Data: Optimal Selling Mechanisms for IoT Data Exchange," *IEEE Transactions on Mobile Computing* 22, no. 4 (2023): 1988–2000.

16. A. Luccioni, S. Viguier, and A. Ligozat, "Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model," *Journal of Machine Learning Research* 24, no. 1 (2023): 1–15.

17. A. Vries, "The Growing Energy Footprint of Artificial Intelligence," *Joule* 7, no. 10 (2023): 2191–2194.

18. R. Buyya, S. Ilager, and P. Arroba, "Energy-Efficiency and Sustainability in New Generation Cloud Computing: A Vision and Directions for Integrated Management of Data Centre Resources and Workloads," *Software: Practice and Experience* 54, no. 1 (2024): 24–38.

19. J. Peng, B. Tang, H. Jiang, et al., "Overcoming Long-Term Catastrophic Forgetting Through Adversarial Neural Pruning and Synaptic Consolidation," *IEEE Transactions on Neural Networks and Learning Systems* 33, no. 9 (2022): 4243–4256.

20. J. Zhou, J. Wan, and F. Zhou, "Transfer Learning Based Long Short-Term Memory Car-Following Model for Adaptive Cruise Control," *IEEE Transactions on Intelligent Transportation Systems* 23, no. 11 (2022): 21345–21359.

21. Y. Fu, C. Liu, D. Li, et al., "Exploring Structural Sparsity of Deep Networks via Inverse Scale Spaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, no. 2 (2023): 1749–1765.

22. S. Wang, Z. Wei, J. Xu, T. Li, and Z. Fan, "Unifying Structure Reasoning and Language Pre-Training for Complex Reasoning Tasks," *IEEE/ACM Transactions on Audio, Speech and Language Processing* 32, no. 1 (2024): 1586–1595.

23. O. Tutsoy, "Pharmacological, Non-Pharmacological Policies and Mutation: An Artificial Intelligence Based Multi-Dimensional Policy Making Algorithm for Controlling the Casualties of the Pandemic Diseases," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, no. 12 (2022): 9477–9488.

24. F. Olan, E. O. Arakpogun, U. Jayawickrama, J. Suklan, and S. Liu, "Sustainable Supply Chain Finance and Supply Networks: The Role of Artificial Intelligence," *IEEE Transactions on Engineering Management* 71 (2024): 13296–13311.

25. Y. Zhang, S. Nie, S. Liang, and W. Liu, "Robust Text Image Recognition via Adversarial Sequence-To-Sequence Domain Adaptation," *IEEE Transactions on Image Processing* 30, no. 1 (2021): 3922–3933.

26. L. Chen, K. Lao, Y. Ma, and Z. Zhang, "Error Modeling and Anomaly Detection of Smart Electricity Meter Using TSVD+L Method," *IEEE Transactions on Instrumentation and Measurement* 71, no. 1 (2022): 1–14.

27. Z. Shi, H. Chen, T. Qu, and S. Yu, "Human–Machine Cooperative Steering Control Considering Mitigating Human–Machine Conflict Based on Driver Trust," *IEEE Transactions on Human-Machine Systems* 52, no. 5 (2022): 1036–1048.

28. J. Berengueres, "How to Regulate Large Language Models for Responsible AI," *IEEE Transactions on Technology and Society* 5, no. 2 (2024): 191–197.

29. G. Maryam and K. Nima, "Ethics in the Age of Algorithms: Unravelling the Impact of Algorithmic Unfairness on Data Analytics Recommendation Acceptance," *Information Systems Journal* (Cambridge, UK: The MIT Press, 2024), 1–32, https://doi.org/10.1111/isj.12572.

30. B. Liu, M. Ding, T. Zhu, Y. Xiang, and W. Zhou, "Adversaries or Allies? Privacy and Deep Learning in Big Data Era," *Information Systems Journal* 31, no. 19 (2024): 31–46.

31. Y. Zhang, W. Gao, H. Yu, J. Dong, and Y. Xia, "Artificial Intelligence-Based Facial Palsy Evaluation: A Survey," *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 31, no. 1 (2024): 3116–3134.

32. S. Kanetkar and V. Thorat, "Ethical Challenges and Guidelines for AI Deployment in Healthcare," *Internet of Medicine for Smart Healthcare* (New Jersey: Wiley, 2024), https://doi.org/10.1002/9781394272266.ch21.

33. J. Zhao, K. Chen, X. Yuan, Y. Qi, W. Zhang, and N. Yu, "Silent Guardian: Protecting Text From Malicious Exploitation by Large Language Models," *IEEE Transactions on Information Forensics and Security* 19 (2024): 8600–8615.

34. D. Kim, Q. Zhu, and H. Eldardiry, "Toward a Policy Approach to Normative Artificial Intelligence Governance: Implications for AI Ethics Education," *IEEE Transactions on Technology and Society* 5, no. 3 (2024): 325–333.

35. N. Kuftinova, A. Ostroukh, O. Maksimychev, A. Podberezkin, and A. Volkov, "Large Language Model in Suburban Transport Data Management," in *2024 Systems of Signals Generating and Processing in the Field of on Board Communications* (Moscow, Russia: IEEE, 2024), 1–5.

36. G. Abich, J. Gava, R. Garibotti, R. Reis, and L. Ost, "Applying Lightweight Soft Error Mitigation Techniques to Embedded Mixed Precision Deep Neural Networks," *IEEE Transactions on Circuits and Systems I: Regular Papers* 68, no. 11 (2021): 4772–4782.

37. H. Shao, Y. Zou, C. Liu, Q. Guo, and D. Zhong, "Learning to Generalize Unseen Dataset for Cross-Dataset Palmprint Recognition," *IEEE Transactions on Information Forensics and Security* 19, no. 1 (2024): 3788–3799.

38. Z. Zhou, M. Shojafar, M. Alazab, and F. Li, "IECL: An Intelligent Energy Consumption Model for Cloud Manufacturing," *IEEE Transactions on Industrial Informatics* 18, no. 12 (2022): 8967–8976.

39. J. Bi, Z. Wang, H. Yuan, J. Zhang, and M. Zhou, "Cost-Minimized Computation Offloading and User Association in Hybrid Cloud and Edge Computing," *IEEE Internet of Things Journal* 11, no. 9 (2024): 16672–16683.

40. R. Sutton, "The Bitter Lesson," 2019.

41. J. Kaplan, S. McCandlish, T. Henighan, et al., "Scaling Laws for Neural Language Models," arXiv:2001.08361v1 .

42. J. Qi and J. Luo, "Small Data Challenges in Big Data Era: A Survey of Recent Progress on Unsupervised and Semi-Supervised Methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, no. 4 (2022): 2168–2187.

43. J. Bi, Z. Wang, H. Yuan, J. Zhang, and M. Zhou, "Self-Adaptive Teaching-Learning-Based Optimizer With Improved RBF and Sparse Autoencoder for High-Dimensional Problems," *Information Sciences* 630, no. 1 (2023): 463–481.

44. Z. Wang, J. Bi, and M. Zhou, "Cost-Minimized Partial Computation Offloading in Heterogeneous Edge-Cloud Systems," *7th International Symposium on Autonomous Systems* (*ISAS*) (2024): 1–6.

45. S. Fu, F. Dong, D. Shen, and T. Lu, "Privacy-Preserving Model Splitting and Quality-Aware Device Association for Federated Edge Learning," *Software: Practice and Experience* 54, no. 10 (2024): 2063–2085.

46. M. Tomei, L. Baraldi, G. Fiameni, S. Bronzin, and R. Cucchiara, "A Computational Approach for Progressive Architecture Shrinkage in Action Recognition," *Software: Practice and Experience* 52, no. 2 (2022): 537–554.

47. J. Fei, Y. Chen, L. Liu, and Y. Fang, "Fuzzy Multiple Hidden Layer Recurrent Neural Control of Nonlinear System Using Terminal Sliding-Mode Controller," *IEEE Transactions on Cybernetics* 52, no. 9 (2022): 9519–9534.

48. D. Chen, X. Li, and S. Li, "A Novel Convolutional Neural Network Model Based on Beetle Antennae Search Optimization Algorithm for Computerized Tomography Diagnosis," *IEEE Transactions on Neural Networks and Learning Systems* 34, no. 3 (2023): 1418–1429.

49. R. Wang, D. Sun, and R. Wong, "Symbolic Minimization on Relational Data," *IEEE Transactions on Knowledge and Data Engineering* 35, no. 9 (2023): 9307–9318.

50. Z. Ji, X. Liu, Y. Pang, W. Ouyang, and X. Li, "Few-Shot Human-Object Interaction Recognition With Semantic-Guided Attentive Prototypes Network," *IEEE Transactions on Image Processing* 30, no. 1 (2021): 1648–1661.

51. S. Teng, J. Li, L. Teng, L. Fei, N. Wu, and W. Zhang, "Scalable Discrete and Asymmetric Unequal Length Hashing Learning for Cross-Modal Retrieval," *IEEE Transactions on Multimedia* 26, no. 1 (2024): 7917–7932.

52. E. Spens and N. Burgess, "A Generative Model of Memory Construction and Consolidation," *Nature Human Behaviour* 8, no. 1 (2024): 526–543.

53. G. Yi and J. Wang, "Frequency-Dependent Energy Demand of Dendritic Responses to Deep Brain Stimulation in Thalamic Neurons: A Model-Based Study," *IEEE Transactions on Neural Networks and Learning Systems* 32, no. 7 (2021): 3056–3068.

54. H. Wu, Z. He, X. Zhang, et al., "ChatEDA: A Large Language Model Powered Autonomous Agent for EDA," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 43, no. 10 (2024): 3184–3197.

55. M. Nashaat and J. Miller, "Towards Efficient Fine-Tuning of Language Models With Organizational Data for Automated Software Review," *IEEE Transactions on Software Engineering* 50, no. 9 (2024): 2240–2253.

56. L. Kloeker, T. Moers, L. Vater, A. Zlocki, and L. Eckstein, "Utilization and Potentials of Unmanned Aerial Vehicles (UAVs) in the Field of Automated Driving: A Survey," *5th International Conference on Vision, Image and Signal Processing* (*ICVISP*). (2021): 9–17.

57. Z. Bai, R. Wang, D. Gao, and X. Chen, "Event Graph Guided Compositional Spatial–Temporal Reasoning for Video Question Answering," *IEEE Transactions on Image Processing* 33, no. 1 (2024): 1109–1121.

58. Q. Shi and L. He, "A Model Predictive Control Approach for Electro-Hydraulic Braking by Wire," *IEEE Transactions on Industrial Informatics* 19, no. 2 (2023): 1380–1388.

59. H. Wang, C. Li, Y. Li, and F. Tsung, "An Intelligent Industrial Visual Monitoring and Maintenance Framework Empowered by Large-Scale Visual and Language Models," *IEEE Transactions on Industrial Cyber-Physical Systems* 2, no. 1 (2024): 166–175.

60. P. Liu, X. Qian, X. Zhao, and B. Tao, "Joint Knowledge Graph and Large Language Model for Fault Diagnosis and Its Application in Aviation Assembly," *IEEE Transactions on Industrial Informatics* 20, no. 6 (2024): 8160–8169.

61. B. Zhou, Z. Yang, Z. Shi, and S. Ma, "Natural Language Processing for Smart Healthcare," *IEEE Reviews in Biomedical Engineering* 17, no. 1 (2024): 4–18.

62. M. A. Akbar, A. A. Khan, S. Mahmood, S. Rafi, and S. Demi, "Trustworthy Artificial Intelligence: A Decision-Making Taxonomy of Potential Challenges," *Software: Practice and Experience* 54, no. 9 (2024): 1621–1650.

63. L. Soni, H. Chandra, and D. S. Gupta, "Post-Quantum Attack Resilience Blockchain-Assisted Data Authentication Protocol for Smart Healthcare System," *Software: Practice and Experience* 54, no. 11 (2024): 2170–2190.

64. J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, and S. Blackwell, “Clinically Applicable Deep Learning for Diagnosis and Referral in Retinal Disease,” *Nature Medicine* 24 (2018): 1342–1350.

65. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, and O. Ronneberger, “Highly Accurate Protein Structure Prediction With AlphaFold,” *Nature* 596 (2021): 583–589.

66. K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, “Accurate Medium-Range Global Weather Forecasting With 3D Neural Networks,” *Nature* 619 (2023): 533–538.

67. J. Lee, W. Yoon, S. Kim, et al., “BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining,” *Bioinformatics* 36, no. 4 (2020): 1234–1240.

68. W. L. Sun and Y. L. Huang, “Cross: A Generic Framework for System Integration and Its Adaption in Hospitals, *Software*,” *Practice and Experience* 52, no. 7 (2022): 1643–1660.