ENGINEERING Artificial Intelligence
https://doi.org/10.2738/ENGAI.2026.0013

Review

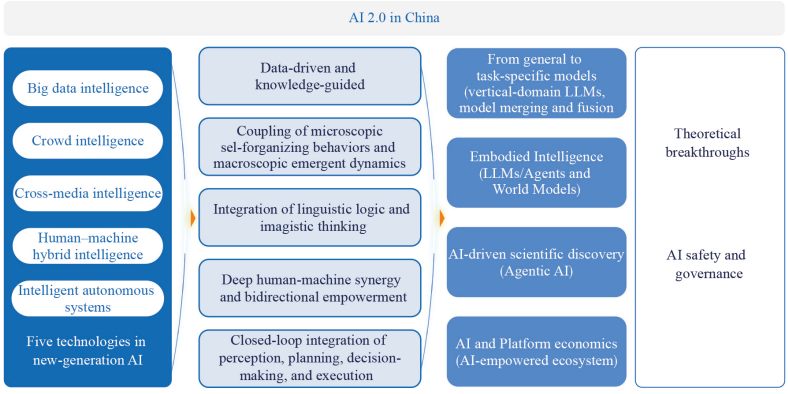
AI 2.0 in China: A Decade of Progress and the Road Ahead

Yunhe Pan (✉)[1], Fei Wu[1], Yi Yang[1], Yongwei Wang[2]

1 College of Artificial Intelligence, Zhejiang University, Hangzhou 310058, China
2 College of Media and International Culture, Zhejiang University, Hangzhou 310058, China

© The Author(s) 2026. This article is published by Higher Education Press.

GRAPHICAL ABSTRACT



HIGHLIGHTS

- Five technological forms define a decade of AI 2.0 in China.
- Collaborative AI, embodied intelligence, and AI-driven discovery highlight current frontiers.
- AI-empowered ecosystems and AI governance shape future directions.

ABSTRACT This paper reviews five technological forms of China's artificial intelligence 2.0 (AI 2.0), namely big-data intelligence, crowd intelligence, cross-media intelligence, human–machine hybrid-augmented intelligence, and intelligent autonomous systems. It further analyzes several major technological trends driven by AI 2.0, including the evolution of large language models from specialized to general-purpose capabilities, embodied agents, AI-driven knowledge discovery, and the AI-driven platform economy. This paper also highlights that the development of novel theories for solving real-world tasks through multidisciplinary integration, along with AI safety and governance, constitutes the major trends in the future evolution of artificial intelligence.

KEYWORDS Artificial intelligence; AI 2.0; Cross-media intelligence; Emerging technologies

1 A Decade Retrospective on the Technological Evolution of AI 2.0 in China

In order to build a first-mover advantage in artificial intelligence (AI) for China, the Chinese Academy of Engineering approved the launch of the major advisory research project "Strategic Research on the Development of China's AI 2.0" in December 2015. The project aimed to examine the significance of AI for China's innovation-driven development and to explore relevant pathways for advancement. The major advisory initiative received strong support from the central government. Under the leadership of the Ministry of Science and Technology and the Chinese Academy of Engineering, the project team conducted in-depth research and completed the draft for the Research Report on Recommendations for the Next Generation Artificial Intelligence Plan.

In July 2017, the State Council issued the Next Generation Artificial Intelligence Development Plan, which stands as China's first national systematic and

Received 24 Aug 2026; accepted 12 Oct 2026

E-mail: pyh@zju.edu.cn (Pan Y.).

strategic development plan for AI in this century. The plan outlines the guiding principles, strategic objectives, key tasks, and supporting measures for China's new-generation AI development up to 2030. It also identified big-data intelligence, crowd intelligence, cross-media intelligence, human–machine hybrid-augmented intelligence, and intelligent autonomous systems as the five technological paradigms of AI. Together, these five paradigms characterize the means and methodologies by which the new generation of artificial intelligence empowers the development of human society [1].

(1) Big-data intelligence. Big data intelligence refers to the theoretical methodologies and supporting technologies that leverage artificial intelligence to conduct in-depth analysis of massive datasets, uncover implicit patterns and regularities, and transform raw data into knowledge, thereby enabling reasoning and decision-making. The current generative large language models, which are built around three core elements of data, models, and computing power, represent a paradigmatic manifestation of the methodology framework of big-data intelligence.

(2) Crowd intelligence. Crowd intelligence refers to the capability of aggregating, organizing, and managing a large number of autonomous participants through specific organizational structures, enabling them to jointly address challenging tasks through diverse autonomous coordination mechanisms such as competition and cooperation. When confronted with complex system tasks in open environments, crowd intelligence can exhibit system-level capabilities beyond those of individual agents. For example, the reasoning and planning capabilities of large models can be combined with the control and action capabilities of agents, enabling embodied intelligent swarms to adaptively accomplish goals and tasks for which no execution paths have been predefined.

(3) Cross-media intelligence. Cross-media intelligence integrates information acquired through multiple perceptual channels, including vision, language, and audition, to construct interoperable semantic representations of the physical world and support functions such as recognition, reasoning, design, creation, and prediction. It serves as a critical information source and enabling mechanism for various intelligent systems to interact with the external environment. Currently, multimodal large language models have found increasingly widespread applications in areas such as human–computer interaction, cross-modal conversion and generation across language, images, audio, and video, as well as robotic control and other related tasks.

(4) Human–machine hybrid-augmented intelligence. Human–machine hybrid-augmented intelligence refers to the integration of human roles or cognitive models into artificial intelligence systems, resulting in an enhanced intelligence paradigm through human–machine synergy. In particular, the "human-in-the-loop" approach has already driven the vigorous development of brain-computer interfaces and wearable intelligent devices.

(5) Intelligent autonomous systems. Intelligent autonomous systems refer to intelligent systems that employ advanced AI technologies to achieve autonomous control and management without direct human intervention. These systems are characterized by capabilities such as autonomous perception, understanding, task planning, decision making, and collaborative coordination. The rapidly emerging fields of embodied intelligence and intelligent agents are representative technological manifestations of this paradigm.

Big-data intelligence extracts patterns and regularities from massive data via the integration of data-driven learning and knowledge guidance. Crowd intelligence emerges from the coupling between microscopic self-organizing behaviors and macroscopic collective dynamics, capturing forms of collective behavioral intelligence observed in biological systems. Cross-media intelligence bridges linguistic reasoning and imagery-based cognition to empower machines with multimodal understanding and generation. Human–machine hybrid-augmented intelligence emphasizes deep collaboration and mutual augmentation between humans and machines, combining the strengths of machine intelligence with those of human cognition to achieve enhanced intelligence. Intelligent autonomous systems establish an integrated perception–planning–decision–execution loop for adaptive action in complex settings. A decade ago, China identified these five paradigms as the technological hallmarks of next-generation AI. Although the development of AI over the past decade has been extraordinarily dynamic and diverse, it has continued to validate the soundness of these five scientific foresights.

The new generation of artificial intelligence features distinctive characteristics of interdisciplinary integration, application-driven empowerment, and a dual nature encompassing both technological and social dimensions. Over the past decade, next-generation AI has advanced rapidly across a wide range of fields, creating substantial opportunities to unleash human creativity and promote economic growth.

Through deep integration with diverse scientific disciplines, AI incorporates domain-specific knowledge into hypothesis generation, algorithm design, optimization and simulation, experimental computation, and mechanism interpretation. This not only significantly accelerates the pace of scientific research but also reveals hidden regularities that traditional methods often fail to

detect, thereby providing new perspectives for scientific breakthroughs and giving rise to a new paradigm of AI-enabled scientific discovery. For example, AI can be used to simulate energy-confinement states under different magnetic-field strengths and plasma densities, rapidly screen potentially optimal experimental parameters, reduce the number and cost of physical experiments, and accelerate research on controlled nuclear fusion. AI can also monitor the operating states and performance of complex engineering systems and provide real-time early warnings of potential security vulnerabilities. In the humanities, AI has been used to digitally reassemble fragments of the Dunhuang manuscripts, restoring the original integrity of ancient texts, gradually piecing together and clarifying cultural relics scattered across the world, and reviving the lost splendor of ancient civilizations [2].

The five technological paradigms of AI 2.0 can be broadly grouped into two categories: "cognitive simulation" and "action simulation". In recent years, technologies such as generative large language models, multimodal large language models, and intelligent content synthesis have continued to advance along the directions of big-data intelligence and cross-media intelligence, demonstrating strong capabilities in tasks such as reasoning in specialized domains. Meanwhile, intelligent autonomous systems, crowd intelligence and human–machine hybrid-augmented intelligence have continued to advance action simulation, empowering unmanned aerial vehicles, autonomous vehicles and vessels, intelligent terminal products, unmanned workshops, ports and mines, as well as integrated human–machine and brain–machine systems. Through "AI+" initiatives, these technologies are reshaping the real economy and accelerating the emergence of a new intelligent economy and intelligent society characterized by human–machine synergy, cross-sector integration, co-creation and sharing. Fig. 1 illustrates the five paradigms with the current and emerging research trends discussed below.

The new generation of artificial intelligence is characterized by a deep coupling of technological and social attributes. The ethical norms it demands have transcended the conventional scope of interpersonal ethics. Rather than concerning only the relationship between humans and given natural facts, AI introduces new forms of interaction and responsibility between humans and human-made artifacts within social systems. The five technological paradigms presented by AI 2.0 involve both fundamental transformations in the production of knowledge and the creation of human-made artifacts in the physical world. As these technologies continue to evolve, human society may enter a new stage characterized by the co-evolution of humans and AI. It is therefore imperative to establish an ethical framework capable of guiding and regulating this evolutionary process and ensuring that it develops in alignment with the shared values of humanity.

This paper analyzes several major technological trends driven by AI 2.0, including the evolution of large language models from specialized to general-purpose systems, embodied agents, AI-driven knowledge discovery, and AI-driven platform economy. It further argues that new theories for solving real-world tasks under interdisciplinary settings, together with AI safety and governance, will constitute important directions for

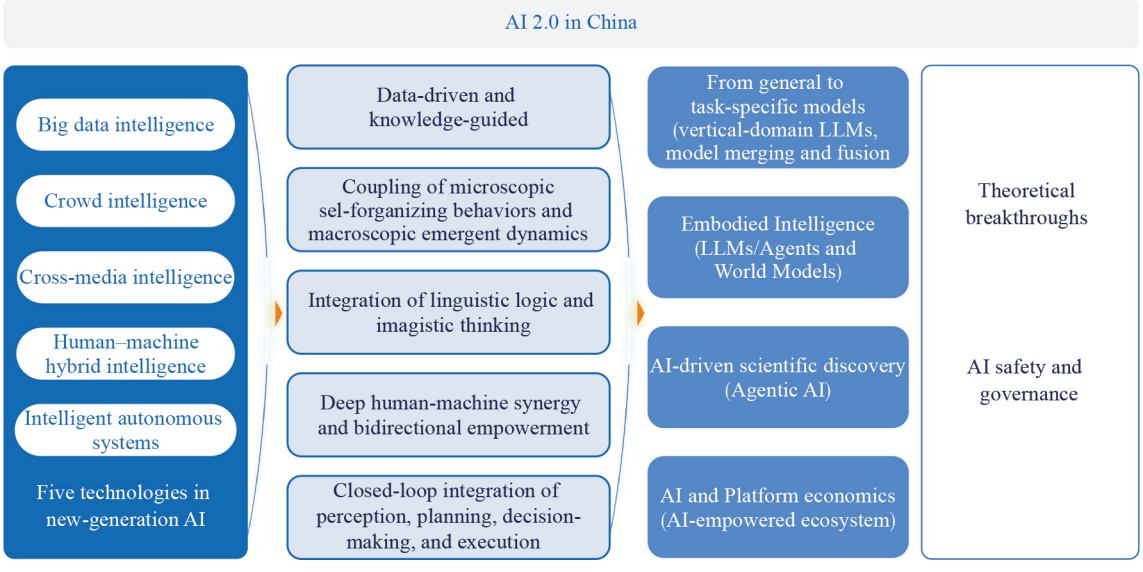


Fig. 1 Five technological paradigms of AI 2.0 and their current research frontiers and emerging trends.

the major trends in the future evolution of artificial intelligence.

2 Emerging Technologies and Research Frontiers in China's AI 2.0

2.1 From specialized to general: the evolving path of large model capabilities

Since the introduction of the Transformer architecture [3], general-purpose large models pre-trained in a self-supervised manner on massive Internet-scale corpora have progressively exhibited emergent reasoning capabilities. Once their parameter scale exceeds a certain threshold, they demonstrate the ability to perform complex reasoning across diverse tasks, instruction following (executing unfamiliar tasks on the fly based on natural language descriptions), and chain-of-thought reasoning (unfolding intermediate reasoning steps for complex problems), while reinforcement learning from human feedback aligns the memory encoded in massive corpora with human value judgments. Building upon this foundational groundwork, and given that large language models can address questions across a wide range of domains, the prospect of artificial general intelligence (AGI) has been reignited. In fact, more than half a century ago, Professor Herbert A. Simon's "General Problem Solver (GPS)" approach similarly inspired hopes for the generalization of AI. Simon's remarkable insight was that he soon recognized the equal importance of knowledge-driven approaches to problem solving. The expert systems developed thereafter precisely combined these two approaches "general reasoning" and "domain knowledge" to address problems in specialized domains.

Contemporary general-purpose large models are facing a similar situation. To efficiently overcome the challenges posed by domain-specific tasks and scenarios, researchers have adopted a "from general to specialized" technical route: first, a general-purpose large model is pre-trained on broad-coverage data and then it is further adapted to specific application scenarios, which is "generalize first, specialize later".

General-purpose large models fundamentally acquire their emergent intelligence from mining the statistical correlations embedded in massive corpora. However, due to a probabilistic-synthesis fallacy, they frequently produce hallucinations in specialized domains. These hallucinations are not random noise, but rather stem from systematic uncertainties arising from insufficient domain-specific knowledge in the training data and from the lack of specialized design and training tailored to professional fields. A 2024 study published in *Nature* [4] proposed a "semantic entropy" measure, which demonstrates that models generate specific types of erroneous responses due to inherent uncertainty, and confirms that such errors exhibit detectable regularities at the semantic level, rather than being purely random noise. To address this limitation, domain-specific large models are typically trained on high-quality in-domain corpora. Taking DeepMind as an example, AlphaFold [5], trained on large-scale domain-specific structural biology data, has raised the accuracy of protein structure prediction to a level approaching that of experimental determination. In mathematics, AlphaProof, reported in *Nature* in 2025, couples a large model with a formal proof-verification environment and trains it through reinforcement learning within a verifiable, domain-specific setting, enabling the system to reach silver-medal-level performance at the International Mathematical Olympiad for the first time [6].

In addition to the "from general to specialized" approach described above, large model capability evolution can also follow a "from specialized to general" pathway (as illustrated in Fig. 2). In this paradigm, a batch of high-quality domain-specific large models is first trained on specialized data drawn from disciplines, industries, and applications, allowing artificial intelligence to first achieve deep and thorough

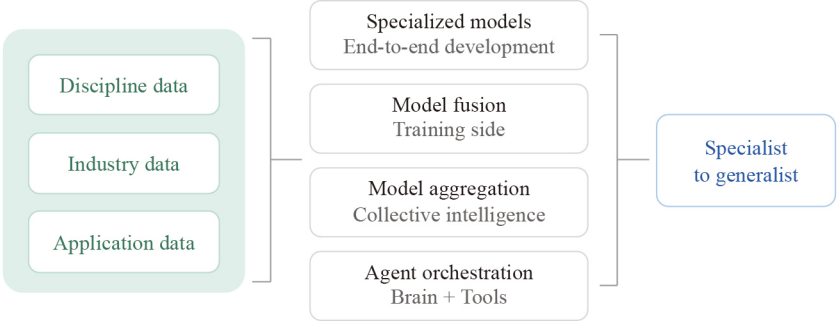


Fig. 2 The specialist-to-generalist path of large-model capability evolution.

proficiency in particular fields. These specialist models are then coordinated with general-purpose capability, complementing each other's strengths and compensating for weaknesses to form a unified, joint intelligent system. In 2025, Zhejiang University advanced this "from specialized to general" route by developing discipline-level and university-level large models. After building the "Zhihai" next-generation AI education and innovation platform, the "Zhihai" series of vertical-domain large models, and the "Mr. ZJU" agent platform, Zhejiang University is building an infrastructure to support large-scale personalized teaching and research. This serves as a concrete implementation of the "from specialized to general" paradigm.

Endowing multiple specialist models with the general capability to accomplish a wide range of tasks does not require that the parameters of all specialist models ultimately be merged into one. As early as 2017, the mixture-of-experts route [7] already demonstrated that combining and invoking multiple sub-networks, each with its own strength, on an as-needed basis can improve a model's overall capability without significantly raising inference cost. This provided the earliest technical prototype for the notion that "a coalition of specialized models can outperform a single generalist model". Since then, model fusion and model aggregation have continued to evolve along distinct technical trajectories. One study [8] showed that an evolutionary algorithm can automatically search for fusion recipes that merge multiple open-source foundation models, each with its own expertise, into a more capable hybrid model without incurring additional training overhead, surpassing even larger single models on tasks such as mathematical reasoning and visual understanding. Another study found that a "Wisdom of the silicon crowd" composed of over a dozen large models can achieve prediction accuracy comparable to that of human prediction groups comprising hundreds of individuals, suggesting that crowd intelligence can also emerge at the model level [9].

The co-scientist system, through mechanisms of generation, reflection, ranking, and evolution, coordinates multiple specialized agents to produce novel research hypotheses for real scientific tasks in biomedicine, mathematics, and physics, some of which have already advanced to experimental validation [10]. In terms of architecture, the "specialized-to-general" route can also take the form of a general-purpose agent that invokes the appropriate specialized model at the right time. HuggingGPT [11], for example, employs a general-purpose model as a task planner that dynamically schedules multiple specialized models to jointly accomplish complex tasks. The cross-fertilization of "specialized" and "generalist" models has already begun to emerge and is becoming an important direction for the next stage of new-generation artificial intelligence.

2.2 Generalization in embodied intelligence

Embodied intelligence represents a crucial form through which artificial intelligence interacts with the physical world. It enables AI to perceive the environment, make autonomous decisions, and execute actions via physical carriers, thereby forming a closed-loop of perception–decision–action in real-world settings [12].

In recent years, embodied intelligence has been advancing toward generalization across tasks and scenarios. RT-1 [13] showed that with large-scale multi-task learning, a single model can learn to handle many different tasks. Open X-Embodiment [14] aggregated data from diverse robotic embodiments, extending scalable learning from a single embodiment to cross-embodiment training, and exploring knowledge transfer across different morphological structures and action spaces. $\pi_{0.5}$ [15] brought the pre-trained knowledge of foundation models into embodied learning, making models better at understanding and carrying out tasks they have never seen before. To address the high cost and limited scale of real-world embodiment interaction data, EgoMimic [16] introduced large-scale human operation data into embodied learning, thereby enriching the training data distribution and improving the model's generalization capability to novel scenarios.

At the same time, the integration of vision-language models and world models has further deepened research in embodied intelligence. Diffusion Policy [17] treats robot control as a problem of generating action sequences, using a diffusion process to capture how multi-step actions relate to one another over time. OpenVLA [18] built action modeling on top of a pre-trained vision-language model, mapping visual observations and language instructions directly to actions. World models, on the other hand, enable embodiments to evaluate potential outcomes of actions before execution by predicting future states under different actions, thereby providing forward-looking information for decision-making [19].

As individual robots become more capable on their own, embodied intelligence is also branching out into collaboration among heterogeneous embodiments and deeper human–machine integration. Different robots have different strengths in perception, movement, and manipulation, and heterogeneous collaboration puts these complementary abilities to work, letting teams of robots take on tasks that no single robot could handle alone [20]. Group collaboration itself is also evolving—from fixed

roles toward dynamic organization, where robots adjust how they divide up and coordinate work as tasks and environments change [21]. Meanwhile, human–machine integration is progressing from the early stage of machines assisting humans in task completion toward closed-loop synergy between human intent and machine perception, decision-making, and execution. Exoskeletons and brain-computer interfaces have emerged as important manifestations of this integration.

Looking ahead, embodied intelligence will need to pay closer attention to safety, practicality, and reliability in the real physical world. On the safety front, embodied agents must be able to recognize and refuse dangerous instructions, while also avoiding unsafe behavior in the course of ordinary tasks—keeping the entire pipeline safe and dependable, from understanding a task to planning and physically executing it. Beyond that, as robot platforms grow more diverse, their differences in form, sensing, and action spaces make it a pressing challenge to transfer knowledge and skills reliably from one embodiment to another. And as models grow larger and tasks more complex, striking a balance between broad capability and real-time responsiveness on resource-limited devices will only become more important.

Although robots represent a paradigmatic technological embodiment of embodied intelligence, the concept of embodied intelligence must transcend the conventional scope of robotics and extend into a broader landscape of machine intelligence, encompassing automobiles, aircraft, ships, machine tools, agricultural machinery, instruments and equipment, among many other forms of machinery. Such an expansion will enable embodied intelligence to better serve humanity across a wider range of real-world applications.

2.3 AI for knowledge discovery

Scientific research constitutes a closed-loop iterative system comprising problem definition, modelling with knowledge, hypothesis generation, experimental validation, result verification, and feedback-driven refinement. In the past ten years, artificial intelligence has empowered individual stages of this loop through specialized models. In modeling with knowledge, DeePMD achieves high-fidelity molecular dynamics simulation through potential energy surface constraints [22], while AlphaFold improves protein structure prediction by leveraging geometric equivariance priors [5]. In experimental validation, Monte Carlo tree search combined with reinforcement learning autonomously plans retrosynthetic routes for target molecules [23], which are later delivered to automated experimental platforms to close the execution loop. Although these models embed local knowledge constraints and perform reliably within their designated settings, the research reliability remains heavily dependent on human intervention: scientists must intervene at critical stages to safeguard scientific rigor, e.g., problem definition, hypothesis selection, data verification, and error correction.

Agentic AI is now systematically displacing this routine human intervention, enabling autonomous iteration across the entire research pipeline. The Co-Scientist [10] and Robin [24] systems, published concurrently in *Nature*, draw on multi-agent division of labor, autonomous decision-making, and dynamic iteration to independently carry out hypothesis generation, experimental design, result interpretation, and solution refinement. Yet the scheduling and reasoning strategies of such agents are learned from static training data and adapt poorly to open, dynamic research environments, so that hypotheses that are internally consistent but fundamentally flawed persist and propagate through successive rounds of iteration.

Verification feedback is therefore emerging as the central bottleneck of this new era of autonomous research. Automated verification fails largely because of cognitive homology: when the verifying agent shares a model backbone, training data, and knowledge base with the generation agent, the two develop substantially overlapping blind spots, so that automated verification can perform only surface-level consistency check and cannot detect deeper errors of scientific mechanism. Effective verification must therefore satisfy two criteria: (1) independence, whereby the basis for verification is exogenous to the model's own generation and inference; and (2) decidability, whereby invalid hypotheses can be precisely rejected without turning to costly wet-lab experiments.

Specialized agents that combine a domain-constrained model with a personalized harness offer a path to satisfying decidability and independence simultaneously, filling in the reliability gap of autonomous research systems. At the model level, a specialized backbone trained on disciplinary mechanisms, industry standards, and domain-specific datasets endows the agent with rigid disciplinary constraints. These specialized backbones are thus capable of identifying invalid conditions, such as out-of-range experimental parameters or unavailable materials, and issuing deterministic rejections. They compensate for the inability to render categorical judgments by general-purpose models, which output only probabilistic confidence scores. At the harness level, a customized layer built from each laboratory's specific equipment, localized experimental protocols, and situated operational experience establishes a verification

dimension external to the model's training regime, breaking the homogeneous blind spots of general-purpose models and providing effective verification on biased autonomous decisions. Therefore, the intelligent scientific research system of the future will not consist of a single general-purpose agent, but rather a cluster of specialized models and intelligent agents that are constrained by domain-specific knowledge, customized for R&D center scenarios, and subject to precise control. This will transform the traditional inefficient model of case-by-case verification into a refined and collaborative workflow based on the classification of scientific research resources. Fig. 3 overviews this paradigm shift in the research workflow.

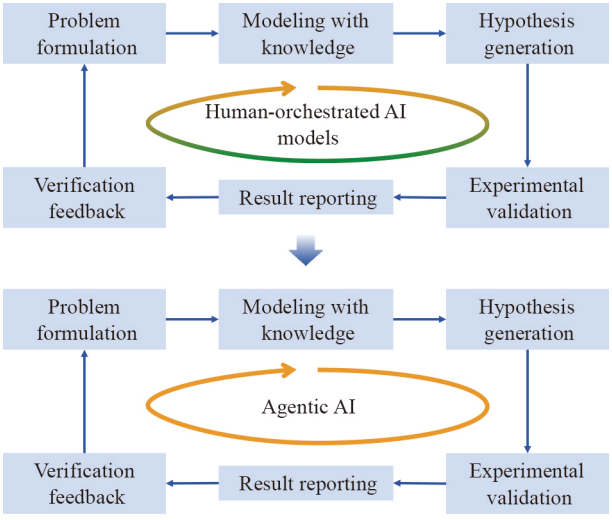


Fig. 3 Illustration of the paradigm shift in scientific research.

2.4 AI-driven platform economy 2.0: from connecting transactions to empowering innovation

The platform economy has emerged as an important outcome of digital technologies reshaping resource allocation and economic organization. Over the past two decades, platforms represented by e-commerce and consumer services have primarily relied on the Internet to connect consumers, businesses, products, and services, with their core value lying in reducing information-search, supply–demand matching, and transaction costs. As AI increasingly permeates economic and social activities, the entities connected by platforms are expanding beyond people, products, and services to encompass data, knowledge, models, tools, and diverse actors across industrial value chains. Accordingly, the mechanisms of value creation are extending from transaction intermediation toward technological capability provision, innovation enablement, and industrial collaboration. We refer to this emerging platform paradigm, characterized primarily by the provision of intelligent capabilities and industrial enablement, as "Platform Economy 2.0".

The fundamental driver of this evolution is the transformation of AI from a collection of task-specific tools into a foundational technology capable of organizing knowledge and capabilities and addressing complex tasks. Advances in big-data intelligence, cross-media intelligence, crowd intelligence, human–machine hybrid-augmented intelligence, and intelligent autonomous systems [1] enable platforms not only to understand and interrelate information from multiple sources, but also to coordinate multiple actors and increasingly support the planning, decision-making, and execution of complex tasks. Platforms are therefore evolving from "connecting information and actors" to "organizing knowledge and capabilities". This transformation has also begun to foster new forms of organization. For example, One-Person Companies (OPCs) supported by large models and intelligent agents are beginning to emerge, reflecting a broader shift in which tasks that traditionally required multiple specialized roles can increasingly be organized and accomplished through platform-based intelligent capabilities and virtual–physical collaborative productive forces.

The essence of Platform Economy 2.0 lies in an AI-driven transition from connection-oriented platforms to capability-oriented platforms, manifested in four dimensions: resource base, platform function, value creation, and competitive advantage. The core resources of platforms are expanding from traffic, users, and transaction data to data, knowledge, models, and specialized tools. Their core functions are evolving from connecting supply and demand and facilitating transactions [25] toward encapsulating complex capabilities into intelligent services that are accessible, replicable, and scalable. Value creation is extending beyond reducing transaction costs, such as those associated with information search and supply–demand matching, to reducing innovation costs, including those associated with knowledge acquisition, access to specialized capabilities, trial-and-error and validation, and organizational coordination. Meanwhile, competitive advantage is shifting from a predominant reliance on user scale and network effects toward a combination of technological capabilities, accumulated knowledge, and the capacity to orchestrate industrial resources.

Reducing innovation costs is therefore a key source of value that distinguishes Platform Economy 2.0 from the traditional platform economy. Traditional Internet platforms primarily improve the allocation efficiency of

existing resources through information aggregation and supply–demand matching. AI-driven platforms go a step further by transforming knowledge, models, and specialized tools into broadly accessible intelligent capabilities, thereby lowering the barriers for individuals and enterprises to engage in innovation activities such as R&D, design, analysis, business operations, and content creation [26]. In this sense, the economic value of platforms is expanding from "making transactions easier" to "making innovation easier". AI+ is thus driving Platform Economy 2.0 beyond the reduction of transaction costs toward the reduction of innovation costs, and beyond the allocation of existing resources toward the creation of new capabilities.

3 Future Directions and Trends

3.1 Fundamental theoretical breakthroughs driven by interdisciplinary research and real-world tasks

A key criterion for assessing whether research constitutes a fundamental theoretical breakthrough is its ability to generate novel knowledge, previously unattainable, through a real, specific, and verifiable task.

Real, verifiable, and interdisciplinary tasks are spurring new theoretical breakthroughs across multiple foundational disciplines. In 2024, the FunSearch combined large models' program-generation capabilities with automated solvers to directly target the long-standing open "cap set" problem in combinatorial mathematics, discovering new mathematical constructions [27]. The following year, a study combining machine learning methods with laboratory-measured three-dimensional particle trajectory data of dusty plasmas demonstrated that long-held theoretical premises in academia—including "particle charge is proportional to radius" and "interparticle forces follow a specific decay law with distance"—lack universal applicability [28]. More recently, a study published in *Nature* combined the chemical requirement of geometric arrangement of catalytic motifs with generative protein structure models to design novel enzymes that have no natural counterparts and cannot be realized by natural proteins [29]. Over the past two years, three foundational disciplines reliant on theoretical deduction—mathematics, physics, and chemistry—have successively yielded examples of new theoretical outcomes driven in reverse by concrete real-world tasks.

A large-scale bibliometric study published in Science in 2013 found that papers with breakthrough impact rarely emerge exclusively from mainstream knowledge combinations within a single discipline. Instead, they often arise from atypical combinations of interdisciplinary knowledge [30]. This aligns with the pattern observed across the three aforementioned cases: breakthroughs in fundamental theory do not necessarily stem from the continuous deduction of an established theoretical system; they may also arise from the impetus of interdisciplinary intersections and real-world tasks.

The mathematical and physical cases discussed above encompass both explorations of theoretical frontiers and machine-verifiable concrete mathematical construction tasks, as well as experimentally measured trajectory data. Real-world tasks furnish abstract theoretical frontiers with tangible targets for empirical testing and falsification. A 2026 perspective article on artificial intelligence-driven scientific discovery [31] similarly notes that amid the widespread integration of generative models into scientific research, verification, alongside traditional hypothesis formulation, is becoming a core bottleneck and central stage in the scientific discovery process. Research paradigms are shifting from a unidirectional "hypothesis–verification" framework to an iterative cycle of alternating generation and verification, a shift that corroborates the logic of using real-world tasks to test theoretical frontiers in reverse.

3.2 AI safety and governance

Over the past decade, AI has evolved from an information-processing tool centered on reasoning and recognition into a complex intelligent system capable of generating knowledge, utilizing tools, and interacting with the physical world, thereby broadening the pathway from cognitive intelligence to behavioral intelligence. Consequently, safety issues are no longer confined to traditional information security concerns such as data leakage and cyberattacks; rather, they have become systemic challenges arising from the interaction among technical capabilities, application scenarios, and social relations.

A core challenge for AI 2.0 safety lies in the structural misalignment between model generation mechanisms and the accountability requirements of an intelligent society. Mainstream models learn statistical correlations from massive volumes of data and generate content in a probabilistic manner, thus they do not inherently possess value objectives or behavioral boundaries aligned with the real world. Biases in training data, insufficient objective specification, and shifts in deployment scenarios may lead to endogenous risks such as hallucinations and bias, while jailbreaking attacks and adversarial examples reveal that model safety constraints can be circumvented by carefully crafted prompts or inputs [32]. Backdoor attacks further indicate that

adversaries may implant anomalous behaviors in advance by poisoning training data or model parameters, to be triggered under specific conditions. As AI agentsbecome connected to retrieval systems, software, and external tools, risks extend from within the model itself to the interfaces through which the system interacts with its environment. Accordingly, the key task of AI safety is not merely to eliminate foreseeable risks, but rather to ensure that intelligent systems remain reliable, auditable, and accountable in complex real-world environments.

Beyond risks endogenous to the model, AI 2.0 also challenges the authenticity of social information and thereby gives rise to derivative risks. AI can rapidly generate multimodal content at scale, including text, images, audio, and video, substantially lowering the cost of producing false information and amplifying its social impact through large-scale dissemination. Current governance efforts mainly follow two technical paths: one is to embed proactive markers such as digital watermarks at the phase of data generation [33]; the other is to accurately identify traces of machine generation at the detection stage using statistical features such as token probability sequences, frequency-domain artifacts, and cross-modal consistency [34]. However, proactive marking depends on deployment by generation platforms and is difficult to extend to open-source large models that do not adopt a unified marking mechanism, as well as to content that has been further processed. In addition, differences across models, domains, and languages, as well as adversarial manipulations such as paraphrasing and compression, can often reduce the generalizability of detectors. Detection techniques therefore provide only probabilistic risk assessments, and the governance of derivative security risks must also rely on coordinated measures, including source authentication, attribution of dissemination responsibility, and platform governance.

A deeper challenge lies in the reshaping of production relations and patterns of social distribution by AI. AI is not merely a source of technological empowerment; it also poses a challenge by driving profound changes in production relations. As intelligent edge devices and AI agents enter production, service, and governance processes, challenging issues will become unavoidable, such as data rights, human–machine collaboration, social distribution, and the protection of workers' rights and interests. AI may not directly replace entire occupations; rather, it is more likely to affect specific tasks across a wide range of occupations [35], thereby raising questions about the revaluation of skills, the reorganization of work processes and collaborative arrangements, and the restructuring of work-related returns. Intelligent education is a representative setting for human–machine collaboration. The *AI Ethics in Education: A Reference Framework* [36], released at the Fourth World Digital Education Conference, puts forward the principles of "human-centered agency, symbiotic collaboration, contextualized beneficence, and differentiated governance", providing an institutional response for clarifying human subject status and the boundaries of responsibility in AI applications. In the era of AI 2.0, the key to safety and governance is to ensure that institutions, ethical norms, and forms of social organization keep pace with the evolution of technology, so that AI can genuinely benefit the development of an intelligent society.

4 Conclusions

A key challenge in advancing the "AI+" industrial empowerment and service transformation lies in how to effectively leverage the generality of existing foundation models alongside the precision of domain-specific knowledge and the rapid transferability of specialized skills. This involves organically integrating physical mechanisms and prior knowledge from professional and industrial domains into deep neural networks; investigating the theories, technologies, and methods for general-specialized integration; establishing an architectural framework for collaborative interaction between general-purpose foundation models and domain-specific expert models; and further exploring technical approaches for the fusion and coordination of models with different specialized capabilities.

The goal is to enable lightweight deployment, knowledge sharing, collective intelligence decision-making, and continuous evolution across diverse expert models and agents in the resolution of complex vertical-domain problems, thereby supporting the construction of intelligent systems that are highly collaborative, lightweight, reliable, and dynamically adaptable [37].

Data availability No datasets were generated or analyzed for this work.

Acknowledgments This research received no external funding. We would like to thank the following scholars (listed in order of the number of strokes in their surnames): Huanghuang Deng, Wenguan Wang, Yueting Zhuang, Zemin Liu, LinchaoZhu, Kun Kuang, Shengyu Zhang, Jun Xiao, Hehe Fan, Zhou Zhao, Ying Wei. The authors used DeepSeek-V4-Pro (online version) as a language-editing tool to improve grammar, clarity, and readability. The AI tool was not used for data analysis, data generation, image processing, or scientific content creation. All outputs were carefully reviewed and verified by the authors, who take full responsibility for the final content.

Competing interests The authors declare no conflicts of interest.

References

1. Pan Y H. Heading toward artificial intelligence 2.0. Engineering, 2016, 2(4): 409−413
2. Zhang Y Q, Fang Z, Yang X Y, et al. Reconnecting the broken civilization: patchwork integration of fragments from ancient manuscripts. In: Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada, 2023, 1157–1166
3. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA, 2017, 6000–6010
4. Farquhar S, Kossen J, Kuhn L, et al. Detecting hallucinations in large language models using semantic entropy. Nature, 2024, 630(8017): 625−630
5. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596(7873): 583−589
6. Hubert T, Mehta R, Sartran L, et al. Olympiad-level formal mathematical reasoning with reinforcement learning. Nature, 2026, 651(8106): 607−613
7. Shazeer N, Mirhoseini A, Maziarz K, et al. Outrageously large neural networks: the sparsely-gated mixture-of-experts layer. In: Proceedings of the 5th International Conference on Learning Representations. Toulon, France, 2017
8. Akiba T, Shing M, Tang Y J, et al. Evolutionary optimization of model merging recipes. Nature Machine Intelligence, 2025, 7(2): 195−204
9. Schoenegger P, Tuminauskaite I, Park P S, et al. Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy. Science Advances, 2024, 10(45): eadp1528
10. Gottweis J, Weng W H, Daryin A, et al. Accelerating scientific discovery with Co-Scientist. Nature, 2026, 655(8122): 487−496
11. Shen Y L, Song K T, Tan X, et al. HuggingGPT: solving AI tasks with ChatGPT and its friends in hugging face. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA, 2023, 1657
12. Liu H P, Guo D, Cangelosi A. Embodied intelligence: a synergy of morphology, action, perception and learning. ACM Computing Surveys, 2025, 57(7): 186
13. Brohan A, Brown N, Carbajal J, et al. RT-1: robotics transformer for real-world control at scale. In: Robotics: Science and Systems 2023. Daegu, Republic of Korea, 2023
14. O'Neill A, Rehman A, Maddukuri A, et al. Open X-embodiment: robotic learning datasets and RT-X models: open X-embodiment collaboration. In: Proceedings of 2024 IEEE International Conference on Robotics and Automation. Yokohama, Japan, 2024, 6892–6903
15. Black K, Brown N, Darpinian J, et al. π0.5: a vision-language-action model with open-world generalization. In: Proceedings ofthe 9th Conference on Robot Learning. Seoul, Republic of Korea, 2025, 17–40
16. Kareer S, Patel D, Punamiya R, et al. EgoMimic: scaling imitation learning via egocentric video. In: Proceedings of 2025 IEEE International Conference on Robotics and Automation. Atlanta, USA, 2025: 13226–13233
17. Chi C, Xu Z J, Feng S Y, et al. Diffusion policy: visuomotor policy learning via action diffusion. The International Journal of Robotics Research, 2025, 44(10-11): 1684−1704
18. Kim M J, Pertsch K, Karamcheti S, et al. OpenVLA: an open-source vision-language-action model. In: Proceedings of the 8th Conference on Robot Learning. Munich, Germany, 2024, 2679–2713
19. Wang H Q, Liang W, Van Gool L, et al. Dreamwalker: mental planning for continuous vision-language navigation. In: Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France, 2023, 10839–10849
20. Cao H Z, Shen J H, Zhang Y, et al. Proximal cooperative aerial manipulation with vertically stacked drones. Nature, 2025, 646(8085): 576−583
21. Li P Y, Zhang S Y, Yang W J, et al. Proactive collaboration via autonomous interaction. Nature Communications, 2026, 17(1): 6306
22. Zhang L F, Han J Q, Wang H, et al. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. Physical Review Letters, 2018, 120(14): 143001
23. Segler M H S, Preuss M, Waller M P. Planning chemical syntheses with deep neural networks and symbolic AI. Nature, 2018, 555(7698): 604−610
24. Ghareeb A E, Chang B, Mitchener L, et al. A multi-agent system for automating scientific discovery. Nature, 2026, 655(8122): 497−505
25. Rochet J C, Tirole J. Platform competition in two-sided markets. Journal of the European Economic Association, 2003, 1(4): 990−1029
26. Noy S, Zhang W. Experimental evidence on the productivity effects of generative artificial intelligence. Science, 2023, 381(6654): 187−192
27. Romera-Paredes B, Barekatain M, Novikov A, et al. Mathematical discoveries from program search with large language models. Nature, 2024, 625(7995): 468−475
28. Yu W T, Abdelaleem E, Nemenman I, et al. Physics-tailored machine learning reveals unexpected physics in dusty plasmas. Proceedings of the National Academy of Sciences of the United States of America, 2025, 122(31): e2505725122
29. Braun M, Tripp A, Chakatok M, et al. Computational enzyme design by catalytic motif scaffolding. Nature, 2026, 649(8095): 237−245
30. Uzzi B, Mukherjee S, Stringer M, et al. Atypical combinations

and scientific impact. Science, 2013, 342(6157): 468−472

31. Cornelio C, Ito T, Cory-Wright R, et al. The need for verification in artificial intelligence-driven scientific discovery. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2026, 384(2317): 20240591
32. Wei A, Haghtalab N, Steinhardt J. Jailbroken: how does LLM safety training fail?. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA, 2023, 3508
33. Dathathri S, See A, Ghaisas S, et al. Scalable watermarking for identifying large language model outputs. Nature, 2024, 634(8035): 818−823
34. Xu Y H, Wang Y W, Bi Y F, et al. Training-free LLM-generated text detection by mining token probability sequences. In: Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore, 2025
35. Eloundou T, Manning S, Mishkin P, et al. GPTs are GPTs: labor market impact potential of LLMs. Science, 2024, 384(6702): 1306−1308
36. Expert Advisory Committee of World Digital Education Alliance, Professional Committee on Ethics and Governance of Artificial Intelligence Open Alliance. Ethics of artificial intelligence in education: a reference framework. Hangzhou: Expert Advisory Committee of World Digital Education Alliance, Professional Committee on Ethics and Governance of Artificial Intelligence Open Alliance, 2026 (in Chinese)
37. Wu F, Shen T, Bäck T, et al. Knowledge-empowered, collaborative, and co-evolving AI models: the post-LLM roadmap. Engineering, 2025, 44: 87−100



Yunhe Pan is a Member of the Chinese Academy of Engineering (CAE) and a Professor at Zhejiang University. He was the former Executive Vice-President of Chinese Academy of Engineering and the former President of Zhejiang University. He is a pioneer in the field of intelligent CAD and computer art in China and has long been working on artificial intelligence, computer graphics, CAD, and industrial design.



Fei Wu received his B.Sc., M.Sc. and Ph.D. degrees in Computer Science from Lanzhou University, University of Macau and Zhejiang University in 1996, 1999 and 2002 respectively. His research interests mainly include Artificial Intelligence, Multimedia Analysis and Retrieval and Machine Learning. From October, 2009 to August 2010, Fei Wu was a visiting scholar at Prof. Bin Yu's group, University of California, Berkeley. Currently, he is a Qiushi Distinguished Professor of Zhejiang University at the College of Artificial Intelligence. He is the dean of undergraduate school of Zhejiang University, and the Executive Director of the National Key Research Base for Textbook (Research on AI Textbooks for Higher Education Institutions).



Yi Yang is a Qiushi Chair Professor at Zhejiang University, and the Director of the Institute of Artificial Intelligence at Zhejiang University. He received the Ph.D. from Zhejiang University in 2010. Prior to joining Zhejiang University, he worked for the University of Queensland, Carnegie Mellon University and the University Technology Sydney. His research is focused on AI technologies and AI deployment. His publication has received over 100,000 citations at Google Scholar. He has received many international and national research awards. Yi Yang is a Fellow of the *IEEE* and the *IAPR*.



Yongwei Wang is a ZJU100 Young Professor at Zhejiang University, and the Vice Director of Institute of Intelligent Audiovisual and Film Arts at Zhejiang University. He received his Ph.D. in Electrical and Computer Engineering from the University of British Columbia, Canada, in 2021, and subsequently worked as a Research Fellow at Nanyang Technological University, Singapore. His research interests include AI for Media (AI4Media), multimedia analysis and forensics, and AI security.