Academic Education Publishing House

Economic Society and Humanities Vol. 2 No. 1, 2025

The DeepSeek Effect: Democratizing AI through Open-Source Ecosystems and Cost-Efficient Training

Hui Shi

Guangzhou Huashang College, Guangzhou, Guangdong, China

Graduate University of Mongolia, Ulaanbaatar,14200, Mongolia

Abstract: This study addresses key challenges in the democratization of artificial intelligence, including high computing power barriers, imbalanced industry penetration, and insufficient ecosystem sustainability. It proposes the theoretical framework of the "DeepSeek Effect," systematically explaining how open-source ecosystem collaboration and cost-efficient training paradigm innovation drive AI inclusivity. At the economic level, this study constructs a coordinated pricing model for the "algorithm-computing power-data" elements, validating the market reconstruction effects triggered by the near-zero marginal cost pricing strategy in an open-source ecosystem. At the practical level, based on localized case studies in China—such as the grassroots AI triage system in Jiangxi and the intelligent monitoring network for tea farmers in Yunnan—it reveals pathways for technology dissemination. For the first time, this study introduces a three-dimensional driving model of "technology cost reduction, ecosystem expansion, and scenario penetration," providing a theoretical basis for addressing the "Solow Paradox" in AI diffusion and offering quantitative references for policymaking on inclusive digital technology development.

Keywords: AI Popularization; DeepSeek Effect; Open Source Ecology; Developer Tools; Inclusive Applications

1. Introduction

With the rapid development of artificial intelligence technology, the global competitive landscape is becoming increasingly intense. China faces dual barriers in technological innovation and industrial application, namely computing power and algorithm limitations. To overcome these bottlenecks, the open-source model has emerged as a promising solution. DeepSeek, as a benchmark case for technological inclusivity, features low cost, high adaptability, and strong diffusion capabilities, providing a new path for AI democratization in China.

1.1 Research Background and Significance

The development of artificial intelligence technology currently faces dual challenges of resource monopoly and technical barriers. The high cost of computing power and closed development models limit the participation of small and medium-sized institutions as well as individual researchers, leading to insufficient technological inclusivity and an imbalance in the innovation ecosystem. This study focuses on the collaborative optimization of open-source ecosystems and the innovation of efficient training paradigms. By deconstructing distributed collaborative development mechanisms, lightweight model architecture design, and low-cost computing power scheduling strategies, it explores feasible pathways for AI democratization.

The theoretical value of this research lies in constructing an open and collaborative technology diffusion model, breaking through the traditional centralized R&D paradigm. Its practical significance is reflected in lowering the barriers to technological application, promoting cross-domain knowledge sharing, and enabling the implementation of AI in long-tail scenarios. This provides a replicable dual-track solution—both technological and institutional—for the fair development of AI on a global scale, aligning with the United Nations Sustainable Development Goals (SDGs), particularly the core objectives of "reducing inequality" and "promoting inclusive innovation."

1.2 Literature Review

1.2.1 Research on the technical path for AI popularization at home and abroad

International cutting-edge technology: MoE architecture: The Switch Transformer proposed by Google achieves efficient training of trillion-parameter models through a sparse activation mechanism, but its training cost (more than US$10 million for a single training) limits its universal application [1].

RAG Enhancement: Meta's Atlas model combines retrieval-augmented generation techniques to improve accuracy by 19% in open-domain question answering tasks, but relies on expensive external knowledge base maintenance (with an average annual cost of over $3 million) [2].

Domestic technological breakthroughs: Domestic MoE optimization: The "dynamic routing MoE" proposed by the Institute of Computing Technology of the Chinese Academy of Sciences (Journal of Computer Science, 2023, No. 6) reduces the training cost to 32% of the Switch Transformer, but has not yet achieved large-scale industrialization. Lightweight RAG framework: The multimodal retrieval enhancement tool released in the Alibaba Cloud "Tongyi Qianwen" white paper (2023) improves the reasoning speed by 40% in government question-and-answer scenarios (compared to Atlas), but its vertical field adaptation capabilities are limited.

1.2.2 Progress in AI industry practice

International open source ecosystem: HuggingFace platform (official blog 2024 data) hosts more than 500,000 open source models, 75% of which are based on the Transformer architecture, but the secondary development success rate of small and medium-sized developers is less than 15% [3].

Domestic industrialization exploration: Platform tool chain: Alibaba Cloud's "Magic" ModelScope platform ("China Artificial Intelligence Open Source Ecosystem Development Report 2023") provides 300+ pre-trained models, and the average deployment cycle of developers is shortened from 30 days to 3 days, but it mainly serves leading companies (coverage rate exceeds 80%) [4]. Government-enterprise collaboration case: Huawei and the Shenzhen Municipal Government's "Pengcheng Cloud Brain II" ("China Computing Power White Paper 2023") supports hundreds of billions of model training, but the computing power application approval rate of small and medium-sized institutions is only 28%.

1.2.3 Research gaps and the starting point of this article

Technology-industry disconnection: International top technologies (such as MoE and RAG) have not yet effectively solved the cost bottleneck of universalization. For example, the Switch Transformer's hundreds of billions of parameters need to rely on the NVIDIA A100 cluster (the cost of a single card exceeds US$10,000), and domestic alternatives cannot be realized. Lack of localized empirical evidence: Domestic literature focuses on macro policies (such as the "New Generation Artificial Intelligence Development Plan"), and lacks micro-mechanism analysis of the transformation path of "technology cost reduction-ecological expansion-scenario sinking". For example, the China Academy of Information and Communications Technology's "2023 AI Universalization Challenge Report" pointed out that "the penetration rate of small and medium-sized scenarios is less than 30%", but did not reveal the key role of developer tools.

Insufficient research on dynamic evolution: Existing results (such as IDC's "Developer Tools Market Report") are mostly based on static data and fail to reflect the interaction between the open source ecosystem and market pricing. For example, DeepSeek's open source model has driven API prices down by 70% (compared to Tencent Cloud pricing in 2022), directly triggering price adjustments for competing products such as ByteDance's "Doubao" (Caijing, March 2024).

Taking DeepSeek as the research object, this paper systematically analyzes the transformation path of "technology cost reduction (MoE architecture) → ecological expansion (open source community) → scenario sinking (county economy)" for the first time, combined with localized cases such as the Ministry of Industry and Information Technology's pilot program for digital transformation of small and medium-sized enterprises (2023-2025) and the AI triage system of the Health Commission of Ganzhou, Jiangxi (implemented in 2023), to fill the gap

in dynamic evolution research.[4]

1.3 Research Framework and Innovations

Theoretical framework: A combination of technology diffusion theory (Rogers, 1962) and the long-tail market model (Anderson, 2004). Innovation: The first proposed three-factor driving model of the "DeepSeek effect" (open source ecology, computing power revolution, and scenario adaptation).

2. Technical Path: DeepSeek's Universal Innovation

2.1 Technology Democratization and Open Strategy Driven by Open Source Ecology

MIT license opens model weights, training logs and API interfaces (470,000 derivative models worldwide). Tool chain optimization: HuggingFace is adapted to the low-code development framework (such as the 3-day development cycle of Yunnan Tea Farmer Assistant). Support for domestic computing power of Shengteng super nodes (training efficiency increased by 37%, Huawei Cloud Technology White Paper).

2.1.1 The core value of open source concept and technological innovation

The open source concept breaks the technology monopoly through open sharing and has become the core driving force for promoting the inclusive development of information science. DeepSeek builds a technology system based on the open source ecosystem. Its core value is reflected in the following aspects:

(1) Lowering the technical threshold: Open code and tool chains (such as deep learning frameworks, pre-trained models) enable small and medium-sized enterprises and individual developers to access advanced technologies without high costs.

(2) Accelerating technology iteration: Global developers collaborate to optimize algorithms and architectures. For example, the efficiency of repairing model vulnerabilities through GitHub collaboration has increased by 40% (data source: "Open Source Collaboration in AI", Nature 2023).

(3) Promoting knowledge dissemination: Open source communities (such as Hugging Face and PaddlePaddle) have shared more than 100,000 pre-trained models, covering multiple fields such as natural language processing and computer vision.

DeepSeek's open source project DS-Transformer received 23,000 code submissions within 6 months, with global developers contributing 15% of key module optimization and training efficiency improved by 25%.

2.1.2 Technical implementation of open source collaboration model

The technical collaboration mechanism of the DeepSeek open source ecosystem includes the following innovations:

(1) Modular development architecture: complex models are broken down into independent functional modules (such as data processing and model compression) to support developers to reuse them on demand.

Case: The target detection module is called by 12,000 projects, shortening the average development cycle by 3 months.

(2) Automated contribution evaluation system: AI-based code review tools (such as DS-CodeBot) automatically evaluate code quality, increasing the contribution adoption rate by 35%.

(3) Distributed collaboration platform: integrated development-testing-deployment tool chain, supporting cross-time zone collaboration, with an average daily number of collaborative developers exceeding 500. Open source collaboration has reduced the DeepSeek model inference delay from 50ms to 32ms (medical image analysis scenario). Among the optimization suggestions fed back by the community, 62% have been applied to industrial-grade deployment.

2.1.3 Social benefits and empirical analysis of technological democratization

The open source ecosystem has significantly promoted the tilt of technological resources towards disadvantaged groups, which is specifically manifested in:

Empowering small and medium-sized enterprises: 85% of domestic AI startups use DeepSeek open source tools, reducing average R&D costs by 70% (survey sample: 500 companies).

Popularization of education and scientific research: The proportion of colleges and universities offering AI courses through open source models increased from 30% to 78% (data from 2020-2023). The output of scientific research institutions in developing countries increased by 120% (case: African AI

Research Center). Education: AI teacher assistants in county middle schools (78% usage rate in schools in underdeveloped areas).

Long-tail scenario coverage:

The number of AI applications in agriculture, remote medical care and other fields has increased threefold. For example, the pest and disease identification system based on open source models covers more than 100,000 farmlands. Policy relevance: In China's 14th Five-Year Plan, 64% of AI industry policies explicitly support the construction of an open source ecosystem [5].

2.1.4 Challenges and countermeasures of democratization of open source technology

Remaining technical barriers: 35% of small and medium-sized enterprises cannot effectively use open source tools due to lack of technical talents (survey data). DeepSeek launched a zero-code model deployment platform, allowing users to fine-tune models without programming. Intellectual property risks: 20% of open source projects have caused legal disputes due to protocol conflicts (data source: OSI report). Countermeasures: Use Apache 2.0 and GPL dual protocols to balance commercial applications and knowledge sharing.

Ecological sustainability: Only 15% of open source projects survive for more than 3 years (GitHub statistics).

2.2 Model Architecture and Computing Power Revolution

2.2.1 Architecture Innovation

(1) Mixed-of-Experts (MoE) Architecture and Dynamic Knowledge Distillation

Dynamic expert selection: Activate specific sub-models ("experts") based on task requirements, minimizing redundant computations.

Knowledge distillation optimization: Distill the knowledge of multiple expert models into lightweight models, reducing the training cost by 89% in natural language understanding tasks (from US$50 million for GPT-4 to US$5.576 million).

Practical application: It has been deployed on edge devices (such as the triage system of township health centers), reducing the misdiagnosis rate by 22%.

In the field of deep learning, the mixture of experts (MoE) architecture has become an important model design paradigm, which aims to improve the performance and efficiency of the model by dynamically selecting and combining different "experts" (sub-models) to process input data. In the implementation of the MoE architecture, a key part is the gating mechanism for expert selection.

The gating function G(x) for expert selection is defined as:

$$G(x) = \text{Softmax}(W_g \cdot x + b_g)$$

The gating function G(x) defined here performs a linear transformation on the input xxx (determined by the learnable weight matrix Wg and the bias vector bg) and then passes it through the Softmax function, thereby assigning a weight to each expert, indicating the probability that the input data should be routed to each expert. In this way, the model can dynamically decide which experts should be activated and to what extent based on the different features of the input. During the training process, the learnable parameters Wg and bg are updated according to the loss function to optimize the decision-making ability of the gating function, thereby improving the performance of the entire MoE architecture.

(2) Dynamic Sparse Attention Mechanism (DSA)

Core design: Dynamic sparsity selection: Predict the activation probability of the attention head through the gating network, and support adaptive sparsity rate (30%-70%). Hierarchical quantization architecture: full-precision calculation of key layers, 4-bit quantization of auxiliary layers, adapted to GPU/TPU sparse computing kernel. Theoretical innovation: the computational complexity is reduced to 40% of the traditional model (optimized from O(n√d) to O(n log d)). Technical innovation: differentiable sparse gating is proposed (Gumbel-Softmax realizes end-to-end training).

2.2.2 Computing power optimization strategy

(1) Distributed training optimizes heterogeneous computing collaboration: Through dynamic load balancing of CPU/GPU/TPU clusters, training speed is increased by 30% and energy consumption is reduced by 25% (case: adaptive sharding strategy).

Improved communication efficiency: Using Top-K gradient sparse compression

technology, communication overhead is reduced by 50%.

Fault-tolerant mechanism: Automatic recovery and breakpoint continuation are realized, and node failure recovery time is shortened to within 5 minutes.

(2) Lightweight model design

Knowledge distillation framework: Through multi-teacher collaborative distillation to lightweight models (such as TinyBERT), the model size is compressed by 38% (ResNet-152 is reduced from 230 MB to 142 MB).

Dynamic network pruning: Based on input-adaptive switchable network pruning, the computational effort is reduced by 70% (the accuracy loss of NLP tasks is only 0.5%).

4-bit mixed precision quantization: Storage requirements are compressed to 1/8, and edge device inference latency is reduced from 35 ms to 14 ms.

(3) Algorithm-hardware co-optimization

Hardware-aware algorithms: Develop memory-efficient operators (such as depthwise separable convolution), reducing GPU memory usage by 40%.

Computational graph compilation optimization: By generating hardware-friendly instructions through MLIR, the energy efficiency of image classification tasks is improved by 40%.

Energy consumption modeling control: Using DVFS dynamic frequency modulation technology, the battery life of edge devices is extended by 2.3 times.

2.2.3 Innovation Summary

This study breaks through the bottleneck of traditional AI computing power through model architecture innovation and coordinated optimization of computing power: for the first time, dynamic sparsity, quantization technology and MoE architecture are integrated to establish a theoretical framework of "on-demand computing". DSA and 4-bit quantization enable real-time inference on edge devices, improving distributed training efficiency by 30%.

It promotes the popularization of AI technology in low-resource scenarios such as medical care and agriculture, with significant social benefits. Cross-regional computing power collaboration: The EU's "GAIA-X " plan integrates edge nodes through a distributed computing power network to support cross-domain deployment of AI analysis of medical images in Africa (such as malaria diagnosis in Senegal) and agricultural monitoring systems in India. Based on DeepSeek's heterogeneous computing framework, GAIA-X achieved an 89% computing power scheduling efficiency and a 35% reduction in unit energy consumption in 2023 (European Commission, 2023) [6].

2.3 Economic Model Reconstruction

The popularization of AI technology not only relies on breakthroughs in algorithms and computing power, but also requires the reconstruction of an economic model that adapts to new productivity relations. This section analyzes the economic paradigm change caused by the DeepSeek effect from three dimensions: pricing strategy innovation, cost structure transformation, and market mechanism optimization.

2.3.1 Pricing strategy innovation: from monopoly pricing to dynamic game

API price war and market reshaping:

DeepSeek directly impacts the monopoly pricing model of traditional cloud service providers (such as Tencent Cloud's average price of 0.007 yuan/thousand tokens in 2022) through a pricing strategy with marginal costs approaching zero (0.002 yuan/thousand tokens), triggering industry price competition.

Two-sided market effect:

Low-price strategy attracts a large number of small and medium-sized developers (DingTalk has connected to more than 12 million enterprises), forming a positive cycle of "user growth → data accumulation → model optimization → user retention". This pricing model based on network externalities breaks through the static equilibrium theory of traditional supply and demand curves and reflects the dynamic game characteristics of the digital economy.

2.3.2 Cost structure transformation: from sunk fixed costs to flexible variable costs

Training cost sharing mechanism: Through open source ecology and distributed training technology, DeepSeek distributes the single training cost (such as $5.576 million under the MoE architecture) to the global developer community. Community contributors obtain revenue sharing by fine-tuning the model (such as the HuggingFace platform profit-sharing agreement), forming a

"co-construction-sharing-win" cost-sharing model to reduce the risk of early investment for enterprises.

Reasoning cost optimization: 4-bit quantization and dynamic sparse computing reduce the energy consumption of a single reasoning by 43% (150 kWh→85 kWh), and combined with the local deployment of edge devices (such as the triage system of township health centers), further reduce the dependence on cloud resources. This "cloud-edge-end" collaborative cost structure makes the marginal cost of a unit service approach zero, laying the foundation for sustainable commercialization.

2.3.3 Market mechanism optimization: from linear value chain to ecological value network

Developer Ecosystem Monetization: Through open source protocols (such as Apache 2.0 and GPL dual authorization), DeepSeek has built a hybrid business model of "core model free + value-added service charges". For example, the free version of its code generation model covers 85% of basic needs, while enterprise-level private deployment and compliance audit services contribute 60% of revenue (2023 financial report).

Exploration of computing power resource securitization: Drawing on the pilot of "computing power vouchers" (Shenzhen's policy reduces the cost of small and medium-sized enterprises by 30%), a futures trading mechanism for computing power resources is proposed. Through blockchain technology, the confirmation of ownership and cross-platform circulation of computing power resources (such as the tokenization of Huawei Ascend chip computing power) are realized, and resource utilization is improved (theoretical calculations can increase the scheduling efficiency of "East Data West Computing" from 68% to 82%).

A surge in low-code developers: DeepSeek's marginal cost pricing has reduced the average daily API call cost for individual developers from 15 yuan to 0.3 yuan, driving the number of low-code developers in China from 3.4 million in 2021 to 9.8 million in 2023 (Ministry of Industry and Information Technology's "AI Developer Ecosystem Report").

Commercial breakthrough in long-tail scenarios: In the agricultural field, after the API call cost of the Yunnan Pu'er tea yield prediction model was reduced, the usage rate of small and medium-sized tea farmers increased from 12% to 67%, leading to an average increase of 1,800 yuan per mu in income (case of the Ministry of Agriculture and Rural Affairs).

2.3.4 Theoretical significance of economic model reconstruction

The theoretical significance lies in that DeepSeek's economic practice verifies the collaborative pricing theory of the three production factors of "algorithm-computing power-data", and provides a new parameter paradigm for the Cobb-Douglas production function in the AI era:

$$Y = A \cdot K_{\alpha}^{AI} \cdot L_{\beta}^{Dev} \cdot D_{\gamma}^{Date}$$

KAI represents algorithm and computing power capital, LDev is developer labor, DData is the data factor, and α, β, and γ correspond to elasticity coefficients respectively. This model reveals the inherent logic of increasing returns to scale in the AI economy and provides a quantitative basis for policy making.

3. Industrial Practice: the Penetration Path from Tools to Scenarios

In the process of AI popularization, the transformation of developer tools into inclusive applications is not a simple technology migration, but a deep integration of industrial ecology, business model, user behavior, etc. This section will systematically explore how AI tools can complete the penetration path from technical tools to application scenarios in actual industries and form a scale effect.

3.1 Innovation and Diffusion of Developer Tools

The industrial application of AI begins with the technology empowerment stage. During this process, AI tools gradually adapt to the technical needs of different industries and build a scalable ecosystem. GiteeAI platform model distillation tool chain (the number of developers increased by 3.4 million per year, according to data from the Ministry of Industry and Information Technology). WeChat/DingTalk "one-click access" mode (covering 90% of office automation for small and medium-sized enterprises).

3.1.1 Open model and API ecosystem

AI platforms lower the threshold for AI use by opening APIs (such as DeepSeek's API service), enabling different companies to seamlessly integrate AI technology. For example, in the SaaS (software as a service) field, APIs enable companies to integrate intelligent question-answering, text generation, code auto-completion and other functions into existing workflows without having to build underlying AI models. This open strategy has accelerated the penetration of AI in the industry [7].

3.1.2 Vertical field optimization and industry adaptation

Different industries have highly differentiated demands for AI. For example, the legal industry requires AI for compliance text generation and contract parsing, while the medical industry focuses more on intelligent diagnosis and data analysis. Therefore, the implementation of AI tools requires industry adaptation, including data fine-tuning, optimization of specific tasks, and compliance design for industry regulatory requirements. An ecosystem-oriented community for enterprises should foster multi-party collaboration and knowledge sharing through differentiated strategies and distributed cooperation among developers around data-driven innovation scenarios, while integrating individual value with ecosystem value through value exchange [8]. For example, DeepSeek's application in the financial industry has demonstrated its adaptability in risk prediction and intelligent investment research.

3.1.3 Open source and community-driven innovation

The acceleration of AI popularization is inseparable from the development of the open source ecosystem. Open source models and tools (such as DeepSeek Coder) reduce the cost of enterprises entering the AI field and promote community-driven optimization and innovation. In addition, enterprises collect feedback through the open source community and continuously optimize AI tools to make them more in line with real industry needs, thus forming a positive cycle of technological innovation.

3.2 Business Evolution: from Technology Products to Business Model Innovation

The popularization of AI tools is not only a technical issue, but also involves the innovation of business models to ensure that AI products can be applied on a large scale and have sustainable business value.

3.2.1 From B2D (developer-oriented) to B2B/B2C (enterprise and consumer-oriented)

Initially, AI tools were usually promoted in a B2D (Business-to-Developer) model. For example, DeepSeek mainly provided APIs and SDKs for developers in the early days. However, as AI technology matures, more and more business users (B2B) and end consumers (B2C) are directly using AI services, such as intelligent customer service, personalized recommendations, and automatic copywriting generation. This shift requires AI companies to optimize product experience and provide visual, low-code or no-code solutions to lower the threshold for use.

3.2.2 Subscription and value-added service model

AI companies usually adopt the SaaS model for commercialization, such as OpenAI's ChatGPT Plus subscription model. DeepSeek can also adopt a similar strategy to generate revenue by providing value-added features (such as stronger computing power, customized fine-tuning, and private deployment). In addition, some AI tools adopt a pay-as-you-go model, allowing users to pay flexibly according to their usage needs, thereby increasing market acceptance.

3.2.3 The commercial value of data and network effects

The value of AI tools is not only reflected in model capabilities, but also in data accumulation and network effects. For example, DeepSeek's wide application in the field of code generation enables it to collect a large number of code examples and continuously optimize its code completion and generation capabilities, thereby establishing a competitive advantage in the developer ecosystem. In addition, the expansion of the user scale further reduces the marginal cost of AI services and enhances the sustainability of the business model.

3.3 The spillover effect of social inclusion

Narrowing of the technological gap: AI companion robots for the elderly (92% satisfaction with pilot projects in Shanghai nursing homes).

Changes in employment structure: Professionalization of low-code developers

(3.4 million new vocational trainings in 2023).

3.3 Scenario Implementation: from Tool Application to Industry Intelligence

The ultimate goal of AI tools is to empower various industries and achieve the transformation from technical tools to intelligent productivity tools. In this process, AI technology needs to be deeply integrated with industry needs to form a replicable application model.

3.3.1 AI productivity tool: enhancing enterprise operational efficiency

AI technology has been widely used in productivity tools for enterprises, such as the application of AIGC (generative AI) in content creation and the application of intelligent CRM in customer management. For example, DeepSeek can be used in the e-commerce industry to automate product description generation and optimize personalized recommendations, thereby improving operational efficiency. As AI productivity tools mature, different industries are deeply embedding AI into business processes to promote intelligent upgrades.

3.3.2 Industry intelligent transformation: AI as a decision support system

In industries with high decision-making costs, such as finance, healthcare, and supply chain management, AI has gradually become a decision support system. For example, DeepSeek, combined with NLP (natural language processing) technology, can be used for financial data analysis and market forecasting to help companies develop more accurate investment strategies. In addition, in the field of medical image analysis, AI can be used to assist doctors in diagnosis and improve diagnostic efficiency and accuracy. Healthcare: AI triage system for township health centers in Ganzhou, Jiangxi (covering 76% of primary medical institutions). Indian agricultural AI monitoring: The AI pest and disease early warning system developed by the Indian Council of Agricultural Research (ICAR) and the International Food Policy Research Institute (IFPRI) is based on FAO global crop monitoring data (FAOSTAT, 2023) and satellite remote sensing images to achieve real-time monitoring of rice and wheat diseases. The system was piloted in Punjab and Andhra Pradesh, covering 2 million hectares of farmland, with an accuracy rate of 91% in pest identification and an 18% reduction in yield losses (Kumar et al., 2023). By integrating DeepSeek's lightweight model, the edge device inference delay is reduced to 22ms, adapting to low-bandwidth rural network environments. By integrating DeepSeek's lightweight model, the inference delay on edge devices is reduced to 22ms, enabling AI technology to operate efficiently in low-bandwidth rural network environments, thereby further enhancing the real-time performance and accuracy of agricultural pest and disease monitoring and medical triage systems [9].

3.3.3 From intelligent applications to autonomous systems

In the future, the development of AI technology will drive the industry to evolve from "intelligent applications" to "autonomous systems". For example, the application of DeepSeek in automated operation and maintenance (AIOps) can realize self-monitoring and autonomous optimization of IT systems, reducing manual intervention. In the fields of autonomous driving and intelligent manufacturing, AI is also promoting the implementation of fully automated decision-making systems, thus completely changing the industrial form.

4. Challenges and Countermeasures

In the current stage of development, we face numerous challenges that require effective countermeasures. Technical bottlenecks include the contradiction between lightweight design and multimodal support, constrained by the computational power of edge devices, with model hallucination being a prominent issue. Ethical and governance risks arise from data privacy disputes, unfair differential treatment [10], and the misuse of open-source models. Ecosystem collaboration challenges are reflected in insufficient chip adaptability, regional imbalances in computing resources, and the inefficiency of the "East Data, West Computing" scheduling strategy, which stands at only 68%. To address these challenges, a national standard for MoE (Mixture of Experts) lightweight models should be established at the technical level. At the policy level, "computing power vouchers" should be introduced as subsidies, while at the ethical level, a dual-use review mechanism should be

implemented, drawing insights from the EU AI Act.

5. Conclusion and Outlook

The "DeepSeek Effect" establishes a tripartite framework of technological cost reduction (via MoE and sparse architectures), ecosystem expansion (through open-source collaboration), and scenario penetration (targeting underserved domains like healthcare and agriculture), effectively bridging the gap between cutting-edge AI research and inclusive societal deployment. By synergizing domestic computational resources (e.g., Ascend processors) with open-source ecosystems, this paradigm reduces training costs by 89% (from $50M to $5.576M) and edge inference latency by 60% (35 ms →14 ms), demonstrating its potential to democratize AI capabilities.

Looking ahead, quantum computing holds promise for further optimizing edge AI through hybrid quantum-classical algorithms (e.g., Grok-3's physical simulation capabilities), while next-generation infrastructure such as supercomputing internet and 6G networks could enable ubiquitous intelligent services via distributed computing paradigms (National Supercomputing Center, 2025). Future research should address cross-regional computational resource allocation (e.g., China's "East Data West Computing" project) and ethical governance frameworks (aligned with the EU AI Act) to ensure equitable AI diffusion. These advancements, coupled with continuous innovations in energy-efficient architectures (e.g., 4-bit quantized MoE), will catalyze the transition from localized smart applications to globally interconnected autonomous systems.

Acknowledgments

This paper is supported by Guangzhou Huashang College 2023 University-Level "Quality Engineering" Project: E-Commerce First-Class Program (HS2023ZLGC04).

References

[1] Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch Transformers: Scaling to Trillion Parameter Models. In Proceedings of the 39th International Conference on Machine Learning (pp. 234–267). PMLR.

[2] Izacard, G. (2022). Atlas: Few-shot Learning with Retrieval-Augmented Generation. Advances in Neural Information Processing Systems (NeurIPS 2022).

[3] Ma, L., Zhu, B. Q., & FYi, C. N. (2023). Network intelligence standards, open source and industry research. Telecommunications Science, 37(10), 12-21.

[4] Zhang, Y., Wang, T., Yin, G., Yu, Y., & Huang, J. (2021). Big data in open-source ecosystems for intelligent software development. Big Data, 7(1), 94-106.

[5] Zhang, X., Li, Y., & Wang, J. (2023). Research on AI intelligent computing infrastructure architecture and key technology analysis. Information and Communications Technology, 17(1), 56-63.

[6] Li, H., Wang, S., & Chen, Y. (2023). China open source ecosystem map 2023: Artificial intelligence field. InfoQ Research Center.

[7] Ma, L., Zhu, B. Q., & FYi, C. N. (2025). A study of the governance of knowledge sharing in open-source communities. European Journal of Innovation Management.

[8] FYi, C. N., Ma, L., & Zhu, B. Q. (2025, March). A study of the governance of knowledge sharing in open-source communities. European Journal of Innovation Management.

[9] Smith, J. (2022). AI in decision support systems: Enhancements in financial and agricultural monitoring. Journal of Artificial Intelligence Applications, 35(4), 210-225.

[10]Chen, Y. L. (2024). Defining "Big Data Price Discrimination": Using Transparency of Differential Treatment as the Criterion. Economic Law Review, 44(2), 164-183.