Logic Mechanism, Practical Dilemmas, and Path Optimization of High-Quality Dataset Construction under Large Model Iteration

Jing Di[a], Na Su[b*], Wei Qin[c]

China Academy of Information and Communications Technology, Beijing, China
[a]dijing@caict.ac.cn, [b*]suna@caict.ac.cn
[c]qinwei@caict.ac.cn

Abstract. Represented by large language models, artificial intelligence technology has entered a new stage characterized by rapid iteration and vertical deployment. As a strategic resource for enhancing model capabilities, the construction level of high-quality datasets directly determines the strength of national digital competitiveness. Breaking through the limitations of existing research that treats datasets and large models as a unidirectional supply-demand relationship, this paper constructs an "empowerment–feedback" analytical framework to systematically demonstrate the bidirectional coupling mechanism between high-quality datasets and large models, and traces the historical evolution and future trends of their adaptive relationship from the perspective of technological history. On this basis, the paper analyzes the structural dilemmas facing China's high-quality dataset construction from four dimensions—foundational resources, quality control, industrial supply, and governance system—and proposes a four-chain synergy optimization path centered on strengthening the data supply chain, enhancing the data value chain, facilitating the data circulation chain, and upgrading the data governance chain.

Keywords: high-quality dataset, large model, bidirectional coupling, data supply, data element governance

1 Introduction

Since the birth of the Transformer architecture in 2017, large language models have demonstrated unprecedented capability leaps in natural language understanding, content generation, logical reasoning, and other domains. From GPT-1's 117 million parameters to GPT-4's trillion-scale Mixture-of-Experts (MoE) architecture, and from single-text generation to multimodal fusion across text, image, audio, and code, large model technology has driven a fundamental paradigm shift in the AI industry from "model-centric" to "data-driven." [1] Data is no longer merely an ancillary input for model training, but a core variable that determines the upper limit of model capabilities—the quantity, quality, and diversity of training data shape, to a considerable extent,

© The Author(s) 2026
B. P. Siba et al. (eds.), *Proceedings of the 2026 5th International Conference on Public Service, Economic Management and Sustainable Development (PESD 2026)*, Advances in Economics, Business and Management Research 411,
https://doi.org/10.2991/978-94-6239-825-2_26

the knowledge boundaries, reasoning depth, and generalization capabilities of large models.

However, the contradiction between the exponential demand of large models for high-quality data and the slowing growth of global data supply is becoming increasingly acute. The Epoch AI research team points out that the stock of high-quality language data will approach depletion around 2026, while low-quality language data will be exhausted between 2030 and 2050 [2]. At the level of Chinese-language data, the situation is even more severe: the White Paper on Training Data for Large Models published by the Alibaba Research Institute reveals that Chinese content accounts for merely 1.3% of the global Internet, compared with 59.8% for English [3]. This scenario not only constrains independent technological innovation, but also concerns the secure transmission of national cultural sovereignty and value orientation in the digital era.

Facing this bottleneck, China has intensively issued policies to promote high-quality dataset construction. In 2022, the Opinions on Building a Data Foundation System to Better Leverage the Role of Data Elements proposed exploring the construction of a data quality standardization system; in December 2024, the Guiding Opinions on Promoting High-Quality Development of the Data Industry first explicitly proposed the concept of "high-quality datasets"; in 2025, the National Data Administration convened a launch meeting for construction work; in June 2026, the Action Implementation Plan for Promoting Industry High-Quality Dataset Construction was issued. By the end of June 2026, China had built more than 120,000 high-quality datasets, with a total volume exceeding 1565 PB [4].

Nevertheless, current academic research on high-quality datasets still predominantly focuses on single dimensions such as evaluation standard system construction, industry scenario application exploration, or construction and operation path design [5]-[6]-[7], treating the relationship between high-quality datasets and large-model construction as a static, unidirectional supply-demand relationship, while overlooking the reverse shaping effect of large-model iteration on dataset construction itself. This paper constructs an "Empowerment–Feedback" analytical framework to reveal the dynamic coupling mechanism between high-quality datasets and large models from the dual perspectives of historical evolution and technological adaptation. On this basis, it systematically analyzes the structural dilemmas confronting China's high-quality dataset construction and proposes a four-chain synergistic optimization strategy.

2 Core Conceptual Definition and Theoretical Positioning

As the core research object of this paper, the conceptual connotation and theoretical positioning of high-quality datasets serve as the logical starting point for subsequent analysis. To accurately grasp the essential characteristics of this concept, it is necessary to trace its evolutionary trajectory and clarify its pivotal role within the theoretical framework of data elements.

2.1 From Dataset to High-Quality Dataset: Conceptual Genealogy and Evolutionary Trajectory

"High-quality dataset" is the result of a genealogical evolution from "dataset" to "high-value dataset" and then to "high-quality dataset," with each conceptual leap reflecting the deepening understanding of data resources within the AI industry's developmental trajectory. Datasets refer to a collection of data organized and stored according to specific structures, formats, or themes. Since the establishment of the UCI Machine Learning Repository by the University of California, Irvine in 1987, datasets have served as the basic unit for machine learning research and experimentation, with their conceptual emphasis lying in fundamentality and instrumentality. High-value datasets emerged from data open governance practices in Europe and America and are defined as data collections capable of generating social benefits and bringing direct returns to users, emphasizing their positive externalities and social attributes. High-quality datasets represent the landmark concept of the current stage of AI development. In December 2024, the Guiding Opinions on Promoting High-Quality Development of the Data Industry first explicitly proposed this concept at the policy level, defining it as "a collection of data that, after processing procedures such as cleaning, annotation, and desensitization, can be directly used for developing and training AI models and effectively improving model performance." Compared with its predecessors, high-quality datasets emphasize adaptability to frontier industries such as AI and large models, as well as intrinsic quality attributes such as completeness, accuracy, consistency, usability, compliance, and non-tampering (see Table 1).

Table 1. Comparison of Datasets, High-Value Datasets, and High-Quality Datasets

Dimension	Dataset	High-Value Dataset	High-quality dataset
Conceptual Definition	A collection of data organized and stored according to specific structures or formats	A collection of data capable of generating social benefits and bringing direct returns to users	A collection of data processed through cleaning, annotation, and desensitization that can be directly used for developing and training AI models and effectively improving model performance
Key Features	Universality, structurality, regularity	High quality, high value-added, high reusability, high efficiency	Well-defined themes, large volumes, diverse categories, up-to-standard quality, high accuracy, and strong readiness for use
Conceptual Emphasis	Foundational, broad concept	Positive externalities, social attributes	Intrinsic quality attributes
Application Scenarios	Widely used in various data analysis scenarios	Data open sharing, data element marketization, cross-domain innovative applications	Large models, AI industry

2.2 The Position of High-Quality Datasets within the Theoretical Framework of Data Elements

The realization of data element value requires a transformation process from data to information, then to knowledge, and ultimately to wisdom, with the prerequisite that data must undergo high-quality processing and standardization. Within the data element theoretical framework, high-quality datasets occupy a critical node connecting “data resourceization” and “data capitalization.” The Guidelines for High-Quality Dataset Construction issued by the National Data Administration in 2025 positions high-quality datasets as the core carrier for integrating AI with the real economy and proposes a three-dimensional classification framework based on data modality, model stage, and industrial application. This implies that high-quality datasets are not merely training materials at the technical level, but a critical nexus connecting the data element institutional system, data infrastructure, and AI industrial applications.

3 Bidirectional Coupling Mechanism between High-Quality Datasets and Large Models

The relationship between high-quality datasets and large models is not a simple "supply–consumption" one, but involves a complex bidirectional coupling mechanism: high-quality datasets empower large models to unleash value, while large-model iteration inversely enhances high-quality dataset construction. This paper summarizes this relationship as the "empowerment–feedback" logic.

3.1 Empowerment: Value Realization of High-Quality Datasets for Large Models

First, capability boundary shaping. High-quality data fundamentally determine the knowledge ceiling and capability boundaries of large models through three functions: description, prescription, and prediction [8]. At the pre-training stage, the structurality and accuracy of data directly influence the model’s depth of language understanding and breadth of reasoning. Compared with unfiltered raw data, high-quality datasets can significantly enhance model robustness and generalizability while reducing overfitting. At the supervised fine-tuning and reinforcement learning stages, high-quality datasets in specific domains inject scenario-based knowledge, guiding large models to construct more targeted and interpretable problem-analysis frameworks. Research indicates that the impact of high-quality data on model performance can be precisely characterized through quantitative approaches: the Quality-aware Scaling Law proposed by Subramanyam et al. introduces the data quality parameter Q into the loss function L(N,D,Q)=A/N<sup>+B/(D<sup> Q<sup>)+E, and systematic experiments demonstrate that when the data corruption rate increases from 0% to 25% (i.e., Q declines from 1.0 to 0.75), the pre-training loss exhibits a predictable, approximately linear increase [9]. This quantitative framework provides mathematical underpinning for the proposition that "data quality determines the upper bound of model performance."

Second, compliance and trustworthiness assurance. Data legality and compliance have become prerequisites for large-model deployment. Once low-quality data contaminated by grammatical errors, false information, content bias, or non-compliant sources are embedded in training corpora, they can easily induce model hallucinations, semantic drift, and privacy infringements. In this sense, the technical risks of large models are essentially externalizations of data risks [10]. Training large models with high-quality datasets characterized by traceability, interpretability, and controllability constitutes a critical step in mitigating technical risks.

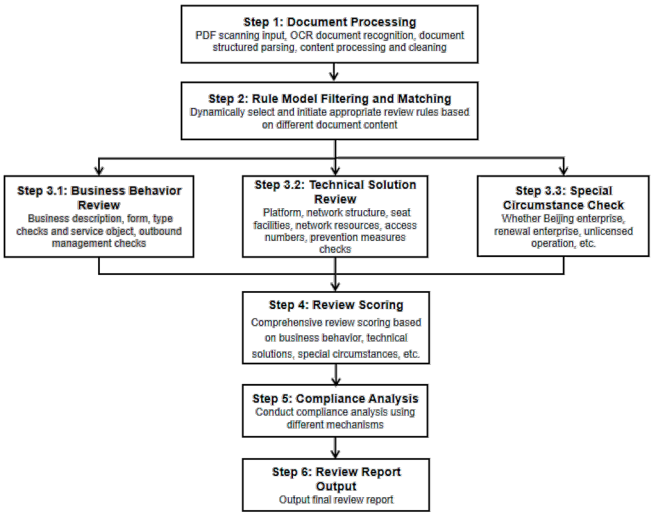


Fig. 1. Intelligent Review Process for Value-Added Telecommunications License Approval

Third, R&D efficiency optimization. High-quality datasets exhibit typical characteristics of externalities and economies of scope [11][12]. By enhancing the semantic accuracy and information density of training samples, high-quality data can substantially compress the parameter scale and iteration cycles required by models, reduce dependence on high-performance computing power, and thereby curtail R&D costs. Meanwhile, if accumulated annotated problematic samples from large-model deployment interactions are fed back and optimized, they can inversely guide the training mechanism of large models and the construction of high-quality datasets. The synergy mechanism formed between the two will facilitate risk control and feedback correction, further improving iteration efficiency. The quantitative correlation between specific data attributes and model behavior has been empirically validated. Regarding data diversity, Bukharin et al., through controlled experiments encompassing over 300 training runs, found that for every 10% improvement in data diversity, the model's average performance on out-of-distribution (OOD) benchmarks increased by approximately 8% [13]. Regarding noise levels, the aforementioned Quality-aware Scaling Law reveals a sublinear decay relationship between the corruption rate and the effective sample size; when the noise proportion remains below 20%, the model demonstrates considerable robustness, with loss growth slower than linear expectations [9]. Regarding domain

coverage, Elazar et al.'s compositional analysis of large-scale corpora demonstrates that the distributional equilibrium of texts from different domains in the training data directly affects the model's generalization capability on downstream tasks, and blind spots in domain coverage often correspond to deficiencies in model capability [14]. Therefore, dataset construction must undergo a paradigm shift from "scale expansion" to "fine-grained attribute control."

The above value-realization mechanisms also have been validated through the intelligent auxiliary approval platform for value-added telecommunications services (see Figure 1) developed by our team. By constructing high-quality industry rule datasets covering 386 core approval rules, the platform transforms scattered policy and regulatory texts into a dual-layer expression structure comprising a "machine-executable layer (YAML/JSON)" and a "semantic interpretation layer (terminology definitions + element checklists + counterexample boundaries)," increasing the semantic coverage rate of rules from 62% under manual collation to 98%, thereby effectively addressing challenges arising from mixed rule granularities and ambiguous semantic boundaries. At the data preprocessing stage, a structure-aware chunking and two-tier semantic retrieval mechanism is adopted: documents are first segmented by block and table cell according to their intermediate representation (Document IR), followed by top-K semantic vector recall based on rule semantics, combined with paragraph-role filtering, ultimately assembling a single-rule context for adjudication. At the semantic understanding stage, element-based semantic adjudication and polarity discrimination subtasks are introduced, enabling the large model to first output an element inspection checklist before aggregating conclusions, effectively resolving issues of fuzzy expressions, near-synonym discrimination, and negation misjudgment. These high-quality dataset construction strategies significantly enhance the review consistency and interpretability of the model in government approval scenarios (see Table 2).

Table 2. Performance Comparison of Algorithm Optimization for Intelligent Auxiliary Review of Value-Added Telecommunications Services

Evaluation Dimension	Typical Pain	Point Algorithm	Improve-ment
Logical Judgment	Directional errors caused by ignoring negation words	Introducing polarity discrimination sub-tasks	Accuracy +17.7%
Semantic Understand-ing	Fuzzy expressions (e.g., "customer communication") prone to misjudgment	Element-based seman-tic adjudication mecha-nism	Precision +25.2%
Review Consistency	Large fluctuations in results across multiple reviews	Deterministic decod-ing and fixed-schema validation	Fluctuation rate -12%
Evidence Recall	Missed recall of key infor-mation in long documents	Structure-aware chunking + two-tier re-trieval	Recall rate +15.7%

3.2 Feedback: Reverse Empowerment of Large-Model Iteration for Dataset Construction

First, data synthesis enhancement. When real-world data are difficult to obtain or excessively costly, leveraging trained large models for data generation and data augmentation can effectively expand dataset scale and richness. In 2025, Microsoft's Phi-4 was trained using no fewer than 50 synthetic datasets, outperforming GPT-4o [15]. In the field of embodied intelligence, a hybrid data system of "1% real-machine + 99% synthetic" is becoming industry consensus, significantly reducing the cost of real-machine data acquisition.

Second, intelligent automated annotation. Through automated and intelligent annotation technologies provided by large models, multi-source heterogeneous data can be standardized and annotated, and customized for specific application scenarios, making the construction of high-quality datasets more efficient. The traditional data annotation industry faces high labor costs, with labor accounting for more than 60% of total project costs [16]. The R&D and application of technologies such as pre-training-assisted annotation, active learning, and weakly supervised learning hold promise for substantially reducing manual annotation costs while improving annotation efficiency and quality.

Third, the data flywheel effect. Large models, through their application of high-quality datasets, feed back into business scenarios and generate new data, thereby producing a "data flywheel" effect—good data trains good models, and good models in turn generate better data, driving continuous improvement in dataset quality. This closed-loop mechanism is particularly evident in application scenarios such as intelligent agents and industry-specific large models: intelligent agents achieve autonomous evolution through a data closed loop of "perception–decision–execution," while industry models enhance scenario adaptability through continuous iteration of the "data–model–scenario" cycle, thereby building industry data moats.

Taking the intelligent auxiliary approval platform for value-added telecommunications services as a further example, the platform also establishes an "approval–feedback–iteration" data closed-loop mechanism. During the review interaction, after the large model generates review conclusions and they undergo human review, error cases and edge cases are automatically fed back to the annotation system, forming incremental training data. Based on a rule pluginization and version management mechanism with hot-loading capability, the platform supports rapid updates during policy iterations; each rule change automatically triggers regression testing and outputs a discrepancy report and risk rule list. The entire review process generates a tamper-proof audit package containing the Document IR, rule hit table, evidence anchors (page number + block_id), model version, and rule version, providing a complete traceable chain for the continuous iteration of dataset quality. The review error rate decreased from an initial 24.3% to 2.8%; each iteration cycle generates approximately 1,400 incremental training samples on average; and the rule library update cycle has been shortened from 15 days to 5 days. This practice demonstrates that interaction data and feedback signals generated by large models in real approval scenario (see Table 3) can effectively feed back into the construction and updating of high-quality datasets.

Table 3. Element Structure of the Review Archival Dataset Generated through Large-Model Feedback

Data Layer	Core Data Elements	Data Format	Business Value
Foundation Layer	Document Intermediate Representation	Structured text (page number, bbox, reading order)	Resolves multi-format compatibility issues for doc/pdf/images; enables cross-page table restoration
Logic Layer	Rule Hit Table	JSON rule index	Enables bidirectional traceability of "rule version + model version"; supports automated regression testing
Evidence Layer	Evidence Anchors	Location information (Block ID), content summary	Provides precise decision-making basis; supports administrative reconsideration and regulatory audit
Governance Layer	Review Archival Package	Complete dataset object	Contains return opinions and rectification lists; feeds back into next-round model training to form a data flywheel

4 Adaptive Logic of the Co-Evolution of Large Models and High-Quality Datasets

The bidirectional coupling mechanism between high-quality datasets and large models (see Figure 2) was gradually formed and continuously deepened over the long course of AI technology evolution. Tracing their co-evolution from the perspective of technological history helps us more profoundly grasp the causes of the current data bottleneck and the direction of future dataset construction.

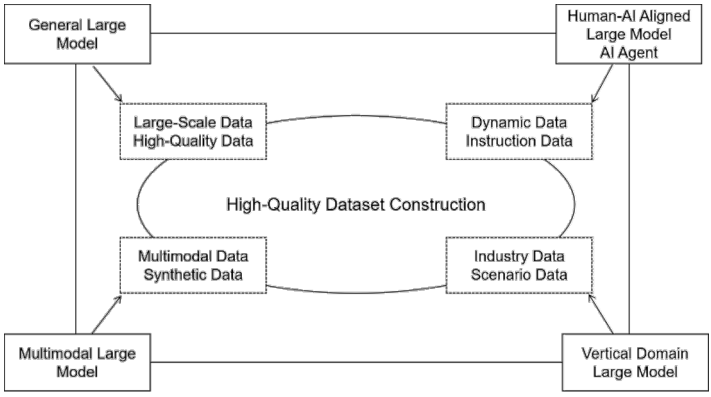


Fig. 2. Adaptive Logic of Large Models and High-Quality Datasets

4.1 Deep Learning Revival and the Rise of Structured Pre-training Datasets (2006–2016)

Although McCulloch and Pitts proposed the mathematical model of neural networks as early as 1943, neural network research progressed slowly for a long time due to limited computing power and data scale. It was not until 2006, when Hinton introduced layer-by-layer unsupervised pre-training to alleviate the vanishing gradient problem, that the concept of neural networks was reactivated [17]. The UCI Machine Learning Repository, established in 1987, was a pioneer in providing structured datasets for machine learning pre-training; the launch of ImageNet by Stanford University in 2009 provided a benchmark testing platform for visual recognition advances; AlexNet's victory in the 2012 ImageNet competition was widely regarded as the dawn of the deep learning revolution.

The core characteristic of this stage was that dataset construction shifted from laboratory manual collation to large-scale Internet acquisition, but data types remained predominantly single-modality structured data. The dataset construction paradigm formed thereby was centered on "category annotation + static storage," with annotation granularity remaining at the coarse-grained label level, data diversity limited by manual design, and no systematic understanding yet formed of the quantitative relationship between data quality and model capabilities. However, the limitations of single-modal structured data became increasingly prominent: visual models required millions of annotated images to converge, while language models were constrained by insufficient knowledge coverage of small-scale annotated corpora. The self-supervised demand of deep learning for large-scale unlabeled text gave rise to a paradigm shift in dataset construction from "manual annotation-driven" to "Internet-scale collection."

4.2 The Transformer Revolution and the Explosion of Large-Scale General Corpora (2017–2021)

The Transformer architecture proposed by the Google team in 2017 fundamentally transformed the technical paradigm of natural language processing. Based on the Transformer architecture and large-scale pre-training corpora, Google and OpenAI successively released billion-parameter pre-trained language models BERT and GPT-1 in 2018, inaugurating the era of large models characterized by big data, big parameters, and big computing power. GPT-2 (2019) and GPT-3 (2020) [18] expanded their parameter scales to 1.5 billion and 175 billion, respectively, with pre-training data requirements reaching 40 GB and 45 TB. Constructing high-quality datasets with high precision, semantic consistency, and refined annotation became the core demand.

The dataset construction paradigm of this stage presented a distinctive "scale-first" characteristic: pre-training corpora leaped from GB-level to TB-level, data diversity was passively guaranteed through large-scale web crawling, and noise control relied on simple heuristic filtering rules. However, systematic attention to the quantitative relationship between data attributes and model behavior remained insufficient, and dataset evaluation still primarily used "scale size" as the main metric. As pre-trained model parameter scales broke through the hundred-billion level, static corpora could no longer

satisfy models' demand for dynamic knowledge updates. The supervised fine-tuning and reinforcement learning in the post-training stage required datasets to shift from "one-time collection" to "continuous iterative updating," making multimodal and interactive data system construction an inevitable choice.

4.3 Generative AI Maturation and the Formation of Multimodal and Dynamic Interactive Data Systems (2022–2025)

The launch of ChatGPT at the end of 2022 marked the entry of generative AI into a mature application stage. Post-training large language models adopted supervised fine-tuning and reinforcement learning paths, breaking through the limitations of using static annotated data. In 2022, massive multimodal datasets such as LAION-5B propelled the maturation of models such as Stable Diffusion and CLIP. Between 2023 and 2024, cross-modal models such as Sora and GPT-4V integrated heterogeneous data into unified systems. In 2025, intelligent agent technologies represented by Manus broke through the limitations of traditional AI applications, making the construction of dynamic interactive datasets a new strategic high ground.

The dataset construction paradigm of this stage underwent a fundamental transformation: from static annotated data to dynamic interactive data, from single-modal independent processing to cross-modal unified alignment, and from the "data flywheel" concept to engineering practice. Datasets are no longer passive training materials, but active carriers of continuous interaction and autonomous evolution between models and environments. After multimodal large models basically established unified heterogeneous data systems, the bottleneck of data supply shifted from "whether there is" to "whether it is good." The maturation of synthetic data technology and the deepening of vertical scenario applications have propelled dataset construction into a new stage of "virtual-real fusion and on-demand customization," where the controllability of data quality and scenario adaptability replace scale as the primary consideration.

4.4 Vertical Scenario-Based and Synthetic Data-Driven Data Paradigms (2026–)

In the future, large models will evolve in parallel along four main trajectories—generalization, multimodality, human-AI alignment, and vertical specialization [19], and the construction of high-quality datasets will accordingly enter a new paradigm phase. First, general-purpose large models will continue to evolve toward greater scale and higher quality, imposing more stringent requirements on both the volume and quality of pre-training data. Second, multimodal large models demand increased cross-modal alignment data, integrating multi-source heterogeneous information encompassing text, images, audio, video, and sensor data. Third, intelligent agents and human-AI alignment-oriented large models necessitate more dynamic interactive data and real-time feedback data, rendering the "dynamicity" and "interactivity" of datasets critical metrics. Fourth, vertical-domain large models call for more professionalized and scenario-specific industry datasets, with the "specialized," "refined," and "precise" attributes of datasets becoming increasingly pronounced. According to Gartner’s prediction, by

2028, the proportion of synthetic data in AI training data will jump from 20% in 2024 to 80%; by 2030 [20], it will surpass real data to become the core basis for business decision-making. With the continuous maturation of frontier generative technologies such as generative adversarial networks, diffusion models, and world models, synthetic data is evolving from a “supplementary resource” to a core component of high-quality datasets, forming a new paradigm of mixed training that integrates “virtual and real” data.

The dataset construction requirements of this stage will revolve around three core dimensions: first, the mixing ratio of synthetic data and real data needs to be dynamically adjusted according to task characteristics, with high-risk fields such as healthcare requiring the real data proportion to be no less than 30%; second, the real-time capture and annotation capability for dynamic interactive data has become a focal point of infrastructure development, requiring interaction logs generated by intelligent agents every second to undergo millisecond-level structured processing; third, the fine-grained construction of vertical scenario data demands deep participation of domain experts in annotation system design, with annotation granularity further sinking from "sentence-level" to "concept-level" and even "reasoning-chain-level." The competitive focus of dataset construction has comprehensively shifted from "data reserves" to "data precision and iteration speed."

5 Practical Dilemmas in China’s High-Quality Dataset Construction

As an important underpinning of AI large models, high-quality datasets have become key elements driving algorithm optimization, model training, and intelligent application deployment. However, in the construction of the data element market and the practice of data governance, China still faces multiple dilemmas in adapting high-quality dataset construction to large-model iteration.

5.1 “Aggregate Poverty” at the Foundational Resource Level

First, the global share disadvantage of Chinese-language corpora. W3Techs data shows that, as of January 2026, English-language websites account for 49.6% of the global total, while Chinese-language websites account for only 1.1%. Looking solely at the pre-training mainstay corpus Common Crawl, the Chinese share is also only approximately 4.8%–5.2%. This forces Chinese large-model enterprises to resort to corpus translation or even extraction from other models to obtain training data, increasing model instability and insecurity.

Second, insufficient openness of public data. As of the end of 2025, among the 26 constituent departments of the State Council, only 17 ministries (65%) have established data openness sections, and they predominantly use non-machine-readable formats such as Word, HTML, and PDF. In contrast, U.S. Data.gov has opened 376,698 datasets in JSON machine-readable format; the EU’s 27 member states average 83% open data maturity.

Third, lagging digitization of original data. National Library statistics show that its electronic books and periodicals exceed 2.2 million titles, but they remain predominantly unstructured scanned documents. China possesses approximately 200,000 existing ancient books, of which fewer than 40,000 have achieved true text digitization, representing a digitization rate of less than 30%. The level of open access to academic resources also shows a significant gap (see Table 4): over the past five years, China's share has remained at 38%–39%, compared with 63% for the U.S. and 49% globally.

Table 4. Global, U.S., and Chinese Open-Access Academic Paper Shares, 2021–2025

Year	Global Open-Access Share	U.S. Open-Access Share	China Open-Access Share
2021	41.0%	55.0%	38.0%
2022	44.0%	59.0%	38.0%
2023	46.0%	61.0%	39.0%
2024	48.0%	62.0%	38.0%
2025	49.0%	63.0%	39.0%

5.2 "Full-Chain Weakness" at the Quality Control Level

The transformation of a high-quality dataset from raw data to a final product typically requires a series of conversion steps. In the data acquisition stage, the Internet remains the primary source of current datasets, but severe homogenization of web data and scarcity of authoritative multimodal data can easily lead to the influx of low-quality text and redundant noise into model training pipelines. In the data cleaning stage, uncleaned raw data are often accompanied by defects such as distribution imbalance and heterogeneous structures; improper cleaning will directly interfere with training effectiveness. In the data annotation stage, 80% of general annotations still rely on manual operations, and the industry lacks a unified annotation quality evaluation system. In the data desensitization stage, when confronted with multimodal data such as documents, tables, images, and videos, existing automated desensitization tools face a dual dilemma of low efficiency and insufficient precision. In the data synthesis stage, data synthesis from multiple sources involves complex issues such as data rights confirmation and privacy protection, bringing legal uncertainty. Synthetic data technology remains at a preliminary stage, with deficiencies in applicability and credibility that can amplify data biases [21]. In the data verification stage, the lack of unified quality evaluation indicators and automated verification tools leads to uneven data quality. The aforementioned stages are interlocking, and weakness in any single stage will be transmitted to downstream stages, forming a cumulative effect of "full-chain weakness".

5.3 "Scenario Disconnection" at the Industry Supply Level

China's current dataset construction faces problems including small data stock and low output, uneven quality, and low utilization efficiency [22]. First, the current data quality standards (GB/T 36344) are oriented toward general information management, without

considering the time-series characteristics, strong correlation, and operational complexity of industrial data; the lack of scenario-based quality grading and collection-annotation specifications renders a significant volume of industrial data difficult to convert into effective training resources. Second, the data trading market is highly fragmented: over 95% of transactions are conducted over-the-counter, with local exchanges adopting inconsistent standards regarding market access, pricing, and compliance, thereby hindering nationwide circulation; simultaneously, the absence of a scientific value evaluation system has resulted in a situation where high-quality data is "priced but unsold", leaving insufficient supply motivation. Third, multi-source cross-domain fusion faces technical obstacles such as format heterogeneity and semantic mapping mismatch, which can easily introduce systematic errors; more critically, there is a lack of trusted security mutual-trust mechanisms and benefit distribution mechanisms, causing enterprises—due to concerns over commercial confidentiality and data security—to generally "dare not share and are unwilling to share," continuously exacerbating the data silo effect.

5.4 "Institutional Lag" at the Governance System Level

First, data ownership systems remain ambiguous: data generation involves multiple stakeholders, and the definition of ownership and division of responsibilities lack a legal basis; in particular, the attribution of new forms such as derived data remains a blank area, making it difficult for market entities to form stable expectations. Second, business models have failed to form positive cycles: the costs of data collection and processing are high, yet differentiated pricing mechanisms and profit models are lacking, resulting in imbalanced input-output ratios; the openness of public data is constrained by security liability pressures and departmental interests, leaving large amounts of high-value data still sealed within public sectors, making it difficult to effectively empower industries. Third, there is a severe shortage of compound talents: the talent gap in the digital economy exceeds 30 million; the talent cultivation system lags behind industry demands in terms of disciplinary settings and industry-education integration; moreover, the lack of unified competency certification standards makes it difficult to support the high-quality construction of data governance systems.

6 Path Optimization for High-Quality Dataset Construction

In the current context of deep integration between the digital economy and artificial intelligence, high-quality datasets are the key lever driving large-model technological innovation and iterative upgrading. In response to the multiple dilemmas facing China's high-quality dataset construction, this paper proposes a four-chain synergy optimization path centered on strengthening the data supply chain, facilitating the data circulation chain, enhancing the data value chain, and upgrading the data governance chain.

6.1 Strengthening the Data Supply Chain: Consolidating the Foundational Base

First, accelerate the opening up of public data resources. At the institutional level, accelerate the implementation of the "Opinions on Accelerating the Development and Utilization of Public Data Resources", clarify the responsibility list and time nodes for data openness of government departments at all levels, strive to achieve full coverage of the data openness section of the State Council's constituent departments, and at the same time build a management and supervision system that covers the entire lifecycle of data, incorporating the quality and efficiency of data openness into government performance evaluation content; At the technical level, we will create a nationwide integrated public data open portal, integrate scattered data resources from various departments and regions, comprehensively promote machine-readable standardized data formats such as JSON, and provide a unified Application Programming Interface (API) to reduce the technical barriers to data acquisition and use; Implement a phased open strategy, prioritize promoting data openness in low sensitive departments, gradually expand to other fields, and establish a high-frequency data update mechanism to ensure data timeliness. Second, lay out the construction of next-generation cutting-edge datasets. Targeting the development needs of next-generation AI technology, we plan to build cutting-edge datasets that are suitable for different training scenarios in advance. For the world model, the focus is on building a multimodal alignment dataset with temporal correlation, covering all scenarios such as static and dynamic, indoor and outdoor, to provide support for the model to establish a unified understanding of the real world; For embodied intelligence, create a hybrid data system of "1% real machine+99% synthesis", integrating real-world robot data with high fidelity simulation synthesis data to reduce the cost of real machine data acquisition; For intelligent agents, build an interactive dataset that is deeply coupled with the testing environment to support the full process training of perception, decision-making, and execution; Based on industry models, we will accumulate professional data in vertical fields such as diagnostic thinking chain, risk control profile, traffic flow, and process knowledge, and create an industry-specific high-quality dataset. At the same time, we will focus on building a diversified thinking chain dataset that covers mathematical reasoning, legal cases, scientific problems, etc., to support large models in breaking through the difficulties of long chain and multimodal reasoning. Third, standardize data supply standards, establish unified data collection formats and metadata specifications for multi-source and multimodal data, and clarify the quality standards for data in the conversion process of "data collection data cleaning data annotation data synthesis productization high-quality dataset". Fourth, activate the conversion of native data value. Accelerate OCR identification and semantic annotation of books, periodicals, archives and other cultural resources, improve the level of open access to academic resources, guide local Internet platforms to establish data exchange mechanisms under the premise of compliance, and break the "data island".

6.2 Enhancing the Data Value Chain: Strengthening Core Competitiveness

First, establish a full lifecycle quality control system. Develop standardized processes covering six stages: collection, cleaning, annotation, desensitization, integration, and validation. Implement a "Data Quality Passport" system, generating unique quality identification codes for each dataset, with shared responsibility between data producers and third-party quality inspection agencies. Deploy automated quality detection pipelines integrating modules such as outlier detection, consistency verification, and semantic drift monitoring, requiring a "dataset quality pass rate" of over 95%. Second, build a professional and intelligent data annotation system. Leveraging the construction of national data annotation bases, enhance precise supply-demand alignment by deeply tapping into specialized annotation needs from libraries, publishers, central enterprises, and large model companies, with a focus on cultivating professional annotation teams in fields like healthcare, industry, and law. Promote intelligent upgrades in annotation technologies, expanding R&D and application of techniques such as pre-trained auxiliary annotation, active learning, and weakly supervised learning to reduce manual annotation costs and improve efficiency and quality. Third, accelerate innovation and application of data synthesis technologies [23]. In terms of technological breakthroughs, increase R&D support for cutting-edge generative technologies like generative adversarial networks, diffusion models, neural radiance fields, and world models, encouraging joint research efforts between academic institutions and enterprises. Construct high-fidelity virtual environments through digital twins and virtual simulations to enhance the physical realism and multimodal alignment accuracy of synthetic data. In terms of industrial ecosystem development, establish a synthetic data standard system covering scenarios such as medical diagnostics, financial risk control, and industrial processes, promoting the adoption of mixed training paradigms combining synthetic and real data, elevating synthetic data from supplementary resources to a core component of high-quality datasets. Fourth, develop a series of national-level industry data spaces. Continuously advance the "AI+" strategy, focusing on key national industries such as manufacturing, healthcare, and transportation, to build national-level industry data spaces that facilitate data sharing and collaborative utilization among upstream and downstream enterprises in the industrial chain. Foster specialized industry data service providers offering end-to-end services including data collection, cleaning, annotation, and modeling, lowering the threshold for SMEs to access industry data.

6.3 Facilitating the Data Circulation Chain: Building a Secure and Efficient Trading System

First, explore the circulation and trading mechanisms for high-value datasets. Establish specialized data service trading platforms, adopting a dual-track model of "standardized exchange on regulated markets + customized agreements off-market," and provide supporting services such as data brokering, compliance assessment, and security auditing. Implement a market-based pricing mechanism under government supervision, taking into account both data quality and application value to form a scientifically rational pricing system. Second, improve the data rights confirmation and benefit distribution

system. Clarify the ownership boundaries and usage standards for public, corporate, and personal data, and establish a flexible benefit allocation mechanism where suppliers and demanders automatically execute rights distribution via smart contracts. Set up data rights registration and dispute arbitration mechanisms to reduce uncertainties in data transactions. Third, enforce compliance and security safeguards. Strengthen the protection system for high-quality datasets, continuously refining the requirements in laws such as the Cybersecurity Law of the People's Republic of China, the Data Security Law of the People's Republic of China, and the Personal Information Protection Law of the People's Republic of China regarding the handling of sensitive data involving personal privacy and trade secrets. Simultaneously, employ technologies like blockchain and privacy computing to fortify data security defenses, enabling free data flow under the premise of "usable yet invisible, controllable yet measurable," thereby fully unleashing the circulation value of data elements.

6.4 Upgrading the Data Governance Chain: Improving Institutional Guarantees and Talent Support

First, improve the data governance system. Refine the legal and regulatory framework by accelerating the introduction of supporting rules in areas such as data ownership, privacy protection, and data security. Clarify the rights, obligations, and responsibility boundaries at each stage of data collection, processing, circulation, and utilization to address core issues like unclear data ownership and ambiguous profit distribution. Establish a classification and tiered protection system for data, implementing differentiated management for public, corporate, and personal data to maximize the flow of data elements while ensuring national security and personal privacy. Build a cross-departmental and cross-regional collaborative governance mechanism to break the deadlock of "multiple authorities with inconsistent standards." Develop a nationally unified data governance framework and quality evaluation system to provide standardized institutional guidance for data supply. Second, enhance the talent cultivation system. Adhere to demand-driven principles, advance the integrated development of education, technology, and talent, and establish a mutually supportive system for data-related disciplines and talent teams across higher education, vocational education, and continuing education. Train interdisciplinary professionals with expertise in artificial intelligence, data engineering, industry applications, and legal compliance. Encourage flexible employment mechanisms such as project collaboration and remote coordination to effectively integrate experienced practitioners from diverse fields and facilitate their deep involvement in data labeling and quality assessment. The National Data Administration, in collaboration with industry associations, leading enterprises, and academic institutions, will jointly establish a unified talent certification system. This system will set evaluation criteria covering technical capabilities, industry knowledge, and compliance literacy for different roles, including data labeling engineers, data quality auditors, and industry data experts, ensuring the certification results are deeply integrated with industry access standards, talent evaluations, and compensation systems.

7 Conclusion

High-quality datasets and large models are dynamic, synergistically evolving inter-constructive entities: high-quality datasets provide core support for large-model iteration through three dimensions of capability boundary shaping, compliance and trustworthiness assurance, and R&D efficiency optimization, while large-model technological evolution inversely drives the scale expansion, quality improvement, and morphological upgrading of high-quality datasets through data synthesis enhancement, intelligent automated annotation, and the data flywheel effect. Their synergistic evolution constitutes the core driving force of AI industry development. The current development trend of large models toward generalization, multimodalization, human-AI alignment, and verticalization determines that high-quality dataset construction must advance toward larger scales, higher quality, multimodal fusion, dynamic interaction, and industry scenario-based orientation. China should strengthen top-level design, improve infrastructure, and drive technological breakthroughs to build an "open, shared, secure, and efficient" high-quality dataset supply system, transforming data resource advantages into AI technology competitiveness and achieving the leap from "following" to "running alongside" and even "leading" in the global AI landscape.

References

1. A. Ng, L. Aroyo, C. Coleman, G. Diamos, V. Reddi, J. Vanschoren, C.-J. Wu, and M. Zaharia, "Data-centric AI workshop," in NeurIPS 2021 Workshop on Data-Centric AI, 2021.
2. P. Villalobos, A. Ho, J. Sevilla, T. Besiroglu, L. Heim, and M. Hobbhahn, "Will we run out of data? Limits of LLM scaling based on human-generated data," arXiv preprint arXiv:2211.04325, 2022. [Online]. Available: https://arxiv.org/abs/2211.04325
3. Alibaba Research Institute, Alibaba Group, Digital China Institute (Fujian), and Alibaba Cloud Intelligence Group, "White paper on training data for large models," Alibaba Research, May 2024. [Online]. Available: https://www.aliyun.com/
4. National Data Administration, "High-quality dataset and data annotation thematic exchange event at the 9th Digital China Summit," National Data Administration, Apr. 2026. [Online]. Available: https://www.nda.gov.cn/
5. Z. Lin, J. Wu, X. Hu, X. Wang, Y. Yuan, and L. Du, "Construction of an evaluation indicator system for high-quality datasets in the data element market," J. Inf. Resources Manage., vol. 15, no. 6, pp. 52–66, 2025.
6. Z. Li, S. Li, and H. Li, "Research on lifecycle model construction of industry high-quality datasets," Intell. Sens. Eng., vol. 3, no. 1, pp. 20–26, 2026.
7. X. Xu, Z. Wu, R. Qiao, A. Verma, Y. Shu, J. Wang et al., "Data-centric AI in the age of large language models," arXiv preprint arXiv:2406.14473, 2024.
8. G. Feng and J. Qiu, "Mechanism and framework of data element value release," Soc. Sci. Shenzhen, vol. 8, no. 1, pp. 47–62, 2025.
9. A. Subramanyam, Y. Chen, and R. L. Grossman, "Scaling laws revisited: Modeling the role of data quality in language model pretraining," in Proc. 14th Int. Conf. Learn. Represent. (ICLR), 2026.
10. L. Zhang, "Accelerate the construction of Chinese training data corpus of AI large models," Frontiers People's Tribune, no. 13, pp. 57–71, 2024.

11. J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark et al., "Training compute-optimal large language models," Adv. Neural Inf. Process. Syst., vol. 35, pp. 14124–14136, 2022.
12. M. Janssen, P. Brous, E. Estevez, L. S. Barbosa, and T. Janowski, "Data governance: Organizing data for trustworthy artificial intelligence," Gov. Inf. Q., vol. 37, no. 3, p. 101493, 2020.
13. A. Bukharin, S. Li, Z. Wang, J. Yang, B. Yin, X. Li, C. Zhang, T. Zhao, and H. Jiang, "Data diversity matters for robust instruction tuning," in Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 3411–3425, 2024.
14. Y. Elazar, A. Bhagia, I. H. Magnusson, A. Ravichander, D. Schwenk, A. Suhr, E. P. Walsh, D. Groeneveld, L. Soldaini, S. Singh, H. Hajishirzi, N. A. Smith, and J. Dodge, "What's in my big data?" in Proc. 12th Int. Conf. Learn. Represent. (ICLR), 2024.
15. Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi, "Self-instruct: Aligning language models with self-generated instructions," in Proc. 61st Annu. Meet. Assoc. Comput. Linguist. (Volume 1: Long Papers), pp. 13484–13508, 2023.
16. B. Ding, C. Qin, L. Liu, Y. K. Chia, B. Li, S. Joty, and L. Bing, "Is GPT-3 a good data annotator?" in Proc. 61st Annu. Meet. Assoc. Comput. Linguist. (Volume 1: Long Papers), pp. 11173–11195, 2023.
17. G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol. 313, no. 5786, pp. 504–507, Jul. 2006.
18. T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., "Language models are few-shot learners," Adv. Neural Inf. Process. Syst., vol. 33, pp. 1877–1901, 2020.
19. W. Fan, S. Li, S. Yan, T. Wang, and F. Cao, "Research on high-quality dataset supply from the perspective of full lifecycle requirements," Inf. Commun. Technol. Policy, vol. 52, no. 5, pp. 2–9, 2026.
20. Y. Huang, S. Wu, C. Gao, D. Chen, Q. Zhang, Y. Wan, T. Zhou, C. Xiao, J. Gao, L. Sun, and X. Zhang, "DataGen: Unified synthetic dataset generation via large language models," in Proc. 13th Int. Conf. Learn. Represent. (ICLR), 2025.
21. Z. Li, H. Zhu, Z. Lu, and M. Yin, "Synthetic data generation with large language models for text classification: Potential and limitations," in Proc. 2023 Conf. Empirical Methods Natural Language Process. (EMNLP), pp. 10443–10461, 2023.
22. L. Long, R. Wang, R. Xiao, J. Zhao, X. Ding, G. Chen, and H. Wang, "On LLMs-driven synthetic data generation, curation, and evaluation: A survey," in Findings of the Association for Computational Linguistics: ACL 2024, pp. 11065–11082, 2024.
23. D. Zha, Z. P. Bhat, K.-H. Lai, F. Yang, Z. Jiang, S. Ding et al., "Data-centric artificial intelligence: A survey," ACM Comput. Surv., 2025, doi: 10.1145/3711118.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

