The Value and Pathways of High-Quality Corpora in Empowering National Intelligent Strategic Communication Capacity*

Han Bo

2026-07-06

ChinaRxiv

Abstract

[Purpose/Significance] The competitive focus of international strategic communication is shifting upstream from the content layer to the foundational layer of training corpora that shape AI cognition. From the perspective of high-quality corpus construction, this study explores its mechanism for empowering national intelligent strategic communication capacity and the practical pathways for shaping intelligent cognitive advantages. [Method/Process] First, the study proposes the concept of "intelligent strategic communication capacity," referring to the ability of actors to shape the internal knowledge frameworks and value logics of artificial intelligence models through the systematic construction of cognitive infrastructures such as corpora. Based on publicly available data on large language model corpora, it comparatively analyzes the linguistic distribution of global corpus resources and the compositional characteristics of training data for mainstream models, assesses China' s cognitive position, and, in combination with policy text analysis, proposes practical pathways. [Results/Conclusions] As a key resource for intelligent strategic communication, corpora have a "double-edged sword" effect: on the one hand, the "dual imbalance" in global corpus resources is leading to systematic intelligent cognitive biases unfavorable to non-Western countries; on the other hand, through strategic construction, they can be transformed into strategic intelligent cognitive advantages. At present, China' s corpus construction still faces problems such as data silos, lack of standards, translation distortion, and the difficulty of enabling high-quality corpora to enter the training processes of global mainstream models. A collaborative governance pathway integrating "values-data-governance" may be explored, providing a reference for connecting the transformation logic from civilizational values to corpus resources and then to strategic communication effectiveness.

Full Text

The Value and Pathways of High-Quality Corpora in Empowering National Intelligent Strategic Communication Capacity*

Han Bo

(Institute of Journalism and Communication, Chinese Academy of Social Sciences, Beijing 100720)

Abstract: [Purpose/Significance] The competitive focus of international strategic communication is moving upstream from the content layer to the foundational layer of training corpora that shape AI cognition. From the perspective of high-quality corpus development, this study explores the mechanisms by which corpora empower national intelligent strategic communication capacity and the practical pathways for shaping advantages in AI cognition.

[**Method/Process**] The study first proposes the concept of "intelligent strategic communication capacity," referring to the ability of actors to shape the internal knowledge framework and value logic of artificial intelligence models through the systematic construction of cognitive infrastructure such as corpora. Based on publicly available data on corpora for large language models, it then comparatively analyzes the linguistic distribution of global corpus resources and the compositional characteristics of training data for mainstream models, evaluates China' s cognitive position, and, in combination with policy-text analysis, proposes practical pathways. [**Results/Conclusion**] As a key resource for intelligent strategic communication, corpora have a "double-edged sword" effect: on the one hand, the "dual imbalance" in global corpus resources is leading to systematic AI cognitive bias unfavorable to non-Western countries; on the other hand, through strategic development, such resources can be transformed into strategic advantages in AI cognition. At present, China' s corpus development still faces problems such as data silos, the absence of standards, distortion in translation and transformation, and the difficulty of integrating high-quality corpora into the training pipelines of mainstream global models. It may therefore be useful to construct a collaborative governance pathway integrating "value—data—governance," providing a reference for connecting the transformation logic from civilizational values to corpus resources and then to strategic communication effectiveness.

Keywords: intelligent strategic communication capacity; corpus; generative artificial intelligence; international strategic communication

Classification number: G206

DOI: 10.31193/SSAP.J.ISSN.2096-6695.2026.02.02

0 Reconstruction of the International Strategic Communication Landscape in the Intelligent Era

In the intelligent era, the way audiences obtain information is shifting from the "search era" to the "question-and-answer era." Artificial intelligence based on large language models continuously intervenes in global users' knowledge acquisition and cognitive construction through everyday question-and-answer interactions[^1], giving rise to a more concealed and more pre-positioned form of strategic communication power. From the perspective of its technical mechanism, this power derives from the model' s statistical learning of massive textual corpora: corpora internalize the associative patterns among vocabulary, grammar, and semantics as the model' s cognitive benchmarks, thereby constituting the foundational infrastructure for shaping AI' s "worldview." The arena of competition in international communication has accordingly shifted: the traditional contest over discourse power focused on "whose voice is louder,"

---

* This article is one of the phased research outcomes of the National Social

Science Fund Youth Project "Research on Pathways for Constructing High-Quality Chinese Media Corpora" (Project No.: 25CXW052).

[**Author Bio**] Han Bo, male, associate research fellow. His research interests include intelligent communication, cyberspace governance, and media economy and management. Email: hanbosmart@outlook.com.

Han Bo. Research on the Value and Pathways of High-Quality Corpora in Empowering National Intelligent Strategic Communication Capacity [J]. *Journal of Documentation and Data*, 2026, 8(2): 016–028.

In the intelligent era, competition has shifted forward to "whose corpus enters AI training data" and "whose narrative framework is learned and reproduced by AI."

Against this background, China's international strategic communication faces new challenges. On the one hand, China's strength in fields related to intelligent technologies is rising rapidly, and the quality and capabilities of domestically developed large models are improving at an accelerating pace. On the other hand, the training corpora of mainstream global models are still dominated by English; high-quality Chinese corpora remain insufficient relative to China's economic scale and cultural resources. When models trained on English-language corpora process issues involving China, they may follow the conceptual frameworks of Western discourse systems and find it difficult to provide accurate interpretations based on native contexts for concepts with Chinese characteristics, thereby producing value bias. In the context in which generative artificial intelligence reshapes knowledge and cognitive production, such "translation distortion" may form intelligent cognitive bias that is both concealed and systematic. Therefore, China's international communication urgently needs a corresponding transformation and upgrading, so as to systematically understand and engage with the laws of intelligent communication in the field of international communication at the level of strategic communication [2].

The analytical focus of traditional strategic communication theory falls on such links as content planning, channel selection, and effect evaluation; it still lacks a new analytical framework for communication phenomena in the intelligent era [3]. On this basis, this study proposes the concept of "intelligent strategic communication capacity," takes corpus construction as the analytical focus, and explores how China can realize the transformation from systematic intelligent cognitive bias to strategic intelligent cognitive advantage, thereby providing reference for enhancing intelligent strategic communication capacity.

1 Intelligent Strategic Communication Capacity: The "Intelligent Turn" of Strategic Communication

As generative artificial intelligence becomes deeply embedded in the global information ecosystem in the role of a "cognitive intermediary," it is necessary to reorganize the practical domains and power configurations of international strategic communication. On the basis of reviewing the theoretical context of strategic communication, this study analyzes the structural challenges that generative artificial intelligence brings to traditional communication paradigms and proposes an analytical framework for "intelligent strategic communication capacity."

1.1 Challenges Posed by Generative Artificial Intelligence to the Paradigm of International Strategic Communication

The classical definition of strategic communication is "the purposeful use of communication by an organization to fulfill its mission" [4]. Its practice and concepts go beyond specific communication "tactics" and pay more attention to the achievement of overall communication objectives at the "strategic" level. Subsequent studies expanded this concept, regarding it as a practice of meaning negotiation and social relationship shaping that is important to the existence of entities [5], and applied it to strategic narrative analysis at the national level [6], exploring how sovereign states shape the international order through narratives. In recent years, the concept of "strategic communication" and its paradigm have gradually been translated and integrated into China's autonomous academic discourse system [2,7], attempting to construct a "multi-subject, three-dimensional communication system" to respond to competition over international discursive power [3].

However, in the face of the rise of generative artificial intelligence, existing research still lacks a relatively systematic theoretical response to the shift in the operational domain of strategic communication from the "content layer" to the "data layer." Traditional strategic communication theory is built on the premise that human actors consciously plan and execute communication; the emergence of generative artificial intelligence has shaken this premise, mainly creating the following challenges.

First, at the level of communication subjects, the clear communication structure presupposed by traditional theory— "communicator-medium-audience" — is becoming blurred in human-machine interaction. As knowledge acquisition shifts from search to prompt-based instant question answering, the subject that generates the answer (such as the developer, corpus source, or algorithm itself) tends to be ambiguous and difficult to attribute clearly. Although model output has formal originality, in essence it remains a reproduction of the statistical patterns of training data. This means that the core assumption of "human

subjects purposefully planning communication" needs to be reexamined. The subject of strategic communication is gradually evolving into a composite subject of "human-algorithm-knowledge," and communication intention and power are generated within a complex network composed of designers, corpus collections, and models' autonomous generation. Second, at the level of mechanisms of power generation, traditional research focuses on the internal

content-layer agenda-setting and frame construction, whereas the cognition of large language models is rooted in statistical learning from training corpora. This means that, by influencing the composition, proportions, and screening of corpora, a more anterior and concealed form of power can operate at the level of the "infrastructure" through which model cognition is formed; its point of action lies closer to the source than traditional content manipulation. Third, at the level of pathways of communication effects, traditional strategic communication models such as soft power and public diplomacy presuppose a relatively open, "explicit" logic of persuasion or attraction, whereas generative artificial intelligence, through routinized and fragmented interaction—continuously providing answers and organizing discourse—imperceptibly shapes users' cognitive frameworks, forming a profound form of "immersive" influence.

When the communicating subject becomes blurred, the generation of power moves upstream, and the pathway of effects shifts toward an immersive state, traditional theories of strategic communication are clearly insufficient for explaining the relationship between communication and power in the intelligent era. New concepts and analytical tools are urgently needed in order to grasp the basic logic of corpora and data as an emerging field of strategic communication.

1.2 Intelligent Strategic Communication Capacity: Concept and Analytical Framework

In the intelligent era, the key to strategic communication is gradually shifting toward shaping the cognitive foundation of content generators. Based on an assessment of how generative artificial intelligence is restructuring the communication paradigm, this study proposes the analytical concept of "intelligent strategic communication capacity," referring to a strategic capacity to define universality through knowledge reshaping. It is the capacity of actors, through the construction of cognitive infrastructures such as corpora and algorithms, to shape the internal knowledge framework and value logic of intelligent models. Centered on the knowledge input and output of intelligent models, this capacity reaches into the sources of AI cognitive formation—data and algorithms—to conduct cognitive calibration and logical embedding; its effects are foundational, concealed, and long-term.

The differences between intelligent strategic communication capacity and traditional strategic communication are mainly reflected on three levels. First, at the level of power generation, traditional strategic communication models place the strategic focus on planning and persuasion at the "content layer," whereas

intelligent strategic communication moves its operation upstream to the "data layer," presetting value frameworks at the source of model cognitive formation by influencing the composition of AI training corpora. Second, at the level of the mechanism of action, intelligent strategic communication is more characterized by "implicit shaping": in the form of infrastructure, it is concealed within everyday human–machine interaction and, through the continuous provision of information and organization of discourse, imperceptibly influences users' cognitive frameworks. This process is often obscured by the "illusion of algorithmic objectivity" [8]. Finally, at the level of the temporal sequence of effects, traditional effects are relatively direct and measurable, whereas the effects of intelligent strategic communication consist in the long-term cumulative reshaping of audiences' cognitive benchmarks regarding what counts as "common sense" and what is "reasonable." The connections and distinctions between intelligent strategic communication capacity and related concepts such as international communication capacity and strategic communication are shown in Table 1.

Table 1 Analysis of Intelligent Strategic Communication Capacity and Related Concepts

Dimension of comparison	International communication capacity	Strategic communication	Intelligent strategic communication capacity
Theoretical definition	The capacity of a country to enhance its visibility and discursive power in the international arena by creating a multi-actor, three-dimensional communication pattern [9–10]	The practical process by which organizations purposefully integrate multiple resources to construct shared meaning and strategic narratives in order to fulfill their missions [3–5]	The capacity formed by actors through the construction of cognitive infrastructures such as corpora to shape the internal knowledge frameworks and value logic of intelligent models; it is a strategic capacity to define universality through knowledge reshaping

Dimension of comparison	International communication capacity	Strategic communication	Intelligent strategic communication capacity
Core object	International audiences, overseas media, etc.	Key audiences and stakeholders	Algorithmic cognitive systems such as artificial intelligence models
Power mechanism	Competition for visibility: emphasizing "reach" in global information flows	Narrative competition: emphasizing identification with meaning and the construction of legitimacy	"Data-algorithm" competition: emphasizing the setting of cognitive benchmarks and logical embedding
Main level of action	Content and channels	Narratives and strategies	Corpora and algorithms
Core logic	Whose voice is heard	Whose narrative is recognized	Whose knowledge and logic are internalized as common sense

To further analyze the generative logic of intelligent strategic communication capacity, this study proposes an analytical framework of "generative foundation—cognitive coordination—generative efficacy"(Table 2). The generative-foundation layer focuses on the source that shapes a model' s value orientation and knowledge structure—the artificial intelligence corpus; the cognitive-coordination layer focuses on the model's ability to transform cognition into content output, acting mainly through intelligent algorithms; and the generative-efficacy layer focuses on how generated content wins users' dependence and trust, ultimately sedimenting into an influence capable of shaping social consensus. The three layers progress in sequence and are interconnected, forming a complete operational chain of intelligent strategic communication capacity.

Table 2 Analytical framework of intelligent strategic communication capacity

Layer	Main sphere of action	Core issue	Analytical dimension	Specific analytical points
Generative foundation	Artificial intelligence corpus	Shaping the model' s knowledge structure and value orientation	Constitutive features	Corpus sources; language/cultural proportions; domain coverage; time span
Generative foundation	Artificial intelligence corpus	Shaping the model' s knowledge structure and value orientation	Quality control	Deduplication; denoising; fact-checking; content-screening standards
Generative foundation	Artificial intelligence corpus	Shaping the model' s knowledge structure and value orientation	Value embedding	Value-annotation system; distribution of narrative frameworks; coverage of cultural concepts
Cognitive coordination	Intelligent algorithms	Influencing users' cognitive framework	Construction of common sense	Presentation of basic facts, historical events, and scientific consensus
Cognitive coordination	Intelligent algorithms	Influencing users' cognitive framework	Framework setting	Ways of defining issues; attribution logic; preferences for solutions
Cognitive coordination	Intelligent algorithms	Influencing users' cognitive framework	Normative shaping	Delimitation of the boundaries of rationality, legitimacy, and acceptability

Layer	Main sphere of action	Core issue	Analytical dimension	Specific analytical points
Generative efficacy	Ideological space	Winning users' dependence and trust	Coverage and reach effects	User scale; interaction frequency; scenario diversity; multilingual support
Generative efficacy	Ideological space	Winning users' dependence and trust	Level of output quality	Factual accuracy; logical coherence; linguistic fluency; safety
Generative efficacy	Ideological space	Winning users' dependence and trust	Iterative matching capability	The degree to which iterative paths such as in-context learning, instruction fine-tuning, and reinforcement learning fit users' cognitive characteristics

Starting from this framework, intelligent strategic communication capacity encompasses multiple dimensions, including algorithms, platforms, interactive scenarios, and training data. Among these many dimensions, the construction of high-quality corpora constitutes the most fundamental and prerequisite link. As the cognitive source and knowledge skeleton of generative artificial intelligence, it plays a foundational role in intelligent communication content and behavior, directly determining the underlying logic by which a model "understands" the world. Therefore, this study focuses on the key dimension of corpus construction, exploring how to systematically intervene in the composition and quality of artificial intelligence training data so as to orderly embed specific value frameworks and narrative logics into the model' s cognitive foundation, thereby continuously

shaping the strategic orientation of its content generation and meaning output.

2 The Mechanism by Which Corpora Empower National Intelligent Strategic Communication Capacity

In the field of international communication, intelligent strategic communication capacity is manifested specifically as national intelligent strategic communication capacity: that is, the deep strategic ability of a state to embed specific narrative coordinates at the cognitive source of artificial intelligence through the construction of high-quality corpora, guide the overall orientation of its global content output, and ultimately influence the international cognitive structure. National intelligent strategic communication capacity is not limited to superficial competition over international discourse power and the sphere of public opinion; more importantly, it concerns the subjective capacity of a country and a civilization in the intelligent era to "define facts,""anchor values,"and "interpret meaning."

Large language models are essentially "stochastic parrots," fluently reproducing the statistical patterns in their training data while being unable truly to understand the meanings behind language

11

. This metaphor indicates that the cognitive capabilities of large language models do not arise out of thin air

arise; rather, it emerges from statistical learning in the corpus. The corpus discussed in this study refers to the large-scale textual datasets used for pretraining large language models; it is the foundational material through which models acquire linguistic rules, factual knowledge, and value orientations. From the perspective of intelligent strategic communication, a corpus is not merely a data collection in the technical sense, but also an infrastructure that shapes a model' s cognitive framework and affects knowledge acquisition by users worldwide. It thus exhibits the characteristics of a cognitive "double-edged sword" : on the one hand, structural defects in corpora may trigger systematic intelligent cognitive bias; on the other hand, through strategic construction and governance, a state can transform corpus resources into a strategic advantage in intelligent cognition.

2.1 Intelligent Cognitive Bias

Intelligent cognitive bias refers to the phenomenon whereby structural defects in the content and sources of corpora cause artificial intelligence systems to form stable and identifiable value biases on particular issues, and to continuously reproduce such biases in large-scale applications. In terms of technical principles, the generation and transmission of this bias are rooted in the core learning mechanism of semantic embedding: the historical and social prejudices

embedded in a corpus are encoded as semantic associations at the underlying level of the model, then converted into generative preferences at the level of probability distributions, and ultimately manifested as value orientations in the output text [12]. Specifically, the formation of intelligent cognitive bias involves three progressive levels.

First, the imbalance in the linguistic distribution of corpora is the starting point of intelligent cognitive bias. Existing large language model corpora exhibit a language-structural bias dominated by English [13]. When Western-language content occupies a dominant position, the modes of expression and conceptual frameworks it contains acquire a "default" status in the semantic structure of the model, causing the model to tend toward reflecting "Western-centric" concepts [14]. Second, the value presuppositions implicit in the corpus enter the model' s cognitive framework through statistical learning. Corpora carry the problem awareness, analytical frameworks, and value positions of their producers. When texts with particular discursive and value characteristics dominate the training corpus, the cognition and values they contain are learned by the model as a "reasonable"linguistic framework. One empirical study compared the answers of five mainstream large language models with country-level data from the World Values Survey and found that mainstream large language models all exhibit cultural value orientations close to those of English-speaking countries and Protestant European countries [15]. This bias is itself the product of a particular cultural context, yet in applications of intelligent technology it is often presented to global audiences under the guise of "objectivity." Finally, conceptual distortion in cross-lingual processing overlaps with the model' s learning mechanism, resulting in the solidification and amplification of bias. Although current large language models pursue multilingual coverage, their original training corpora are still dominated by English and other leading languages. The model' s understanding of concepts in other languages often has to rely on semantic association networks formed within English corpora. This means that once indigenous concepts enter the cross-lingual processing workflow, they are forcibly embedded in a heterogeneous semantic coordinate system. Meanwhile, narratives that appear frequently in the corpus and are supported by complete argumentation are assigned higher probability weights during generation. Once a particular narrative occupies a dominant position in the training corpus, it creates a self-reinforcing cycle of bias [16].

This bias is especially prominent when concepts related to China are involved. When models trained on English corpora process Chinese or China-related issues, they often need to map Chinese concepts onto approximate concepts already defined in English corpora. This black-box process of "conceptual mapping" can easily lead to distortion in cross-cultural translation and imbalance in values. Taking concepts related to traditional Chinese medicine as an example, studies have found that, because the proportion of Chinese-language content in the training corpora of Western models is relatively low, such models have difficulty accurately understanding particular Chinese concepts and terminological systems [17]. Multilingualism is not the same as multiculturalism. When a

model lacks sufficient support from local corpora, it may simply map concepts carrying specific cultural connotations onto approximate terms in a Western context, causing bias in analytical frameworks and evaluative standards. Models need exposure to culture-specific content in order to develop cross-cultural understanding [18].

2.2 Intelligent Cognitive Advantage

In contrast to intelligent cognitive bias, intelligent cognitive advantage refers to the ability of an actor, through proactive planning and the systematic construction of high-quality corpora, to implant favorable value frameworks and knowledge structures into the cognitive foundation layer of artificial intelligence, thereby gaining sustained influence in national intelligent strategic communication

ability. This advantage can be understood in terms of three interrelated dimensions: knowledge supply, value anchoring, and cultural interpretation.

Knowledge supply provides a solid and reliable knowledge foundation for model construction. The "hallucination" phenomenon in large language models is, in essence, a product of insufficient knowledge supply. When a model lacks adequate training data on a specific issue, it tends to fill cognitive gaps with statistically "reasonable guesses." To suppress "hallucinations" and cognitive blind spots caused by corpus scarcity, it is necessary to provide corpora that are sufficient in scale, balanced across domains, and factually accurate, covering multiple dimensions such as politics, the economy, culture, and science and technology, so that the model possesses systematic and trustworthy knowledge reserves on relevant issues. On this basis, model outputs can become more fact-based and persuasive, thereby performing more reliably on issues involving China, rather than producing biased "reasonable fabrications."

Value anchoring embeds in the model a coordinate benchmark for the rationality of judgment. The output of a large language model is not only a presentation of facts, but also a product of value judgment. When generating text, the model needs to make implicit judgments about the "importance" and "reasonableness" of information, and the benchmark for such judgments is learned precisely from the value distribution of the training corpus. Therefore, the systematic supply of high-quality corpora with clear value orientations can influence the model' s internal benchmark of rationality. Studies have found that prompt engineering that explicitly specifies cultural identity can significantly improve the degree of alignment between model outputs and the cultural values of a particular country [15], indicating that data intervention can effectively adjust the model' s mechanism of value generation. For China, if high-quality Chinese corpora can effectively enter the training process of mainstream large language models, they will help establish, at the cognitive foundation of the model, a framework of value judgment adapted to the Chinese context.

Cultural interpretation ensures the semantic integrity of key concepts in cross-

linguistic mapping. Conceptual distortion in cross-linguistic processing is rooted in the model' s insufficient exposure to the native contexts of local concepts. To prevent emblematic concepts such as "Chinese-style modernization" and "whole-process people's democracy"from being simplified, mistranslated, or even distorted in cross-linguistic output, it is necessary to provide a large volume of native local corpora with complete contexts and clear explanations, enabling the model to learn the semantic networks and cultural boundaries of concepts from actual usage. Only when the model forms an independent and complete conceptual understanding on the basis of sufficient "culture-specific content" can it restore the original meaning of Chinese narratives in global communication and form correct Chinese expression and Chinese cognition.

The three dimensions of knowledge supply, value anchoring, and cultural interpretation are interconnected and together constitute the basic logic by which corpus resources are transformed into a strategic advantage in intelligent cognition. In practice, they point to the basic characteristics of what constitutes a "high-quality" corpus in the context of strategic communication. Factual accuracy is the foundation of knowledge supply, ensuring that model outputs are persuasive; value orientation is the goal of value anchoring, enabling model judgments to conform to strategic interests; and cultural integrity is the guarantee of cultural interpretation, preserving the native meaning of core concepts in cross-linguistic communication.

3 The Global Corpus-Construction Landscape and China' s Cognitive Positioning

The discussion of the strategic value of corpora must be grounded in a systematic assessment of the global resource landscape and competitive dynamics. To objectively analyze China' s cognitive positioning in the era of intelligent communication, this study examines the construction landscape of global large language model corpora from three levels: resource distribution, model training, and infrastructure development.

3.1 Current Distribution of Global Corpus Resources

At present, the global distribution of corpus resources generally exhibits a "dual" imbalance.

The first imbalance is the mismatch between language population and the share of digital content—that is, the proportion of a given language in global website content (Table 3). The number of global Chinese users and English users, including both native and second-language users, is roughly comparable [19], whereas the language of Internet content···

the language shares are imbalanced. According to W3Techs monitoring data for the world' s top 10 million websites by traffic, as of December 12, 2025, English content accounted for as much as 49.6%, while Chinese content accounted for

only 1.3% [20].

Table 3 Population size of major world languages and their share of website content

Language	Users (100 million)	Share of world population	Share of website content	Digital representativeness coefficient
English	15	18.8%	49.6%	2.65
Spanish	5.6	7.0%	6.0%	0.86
German	1.3	1.6%	6.9%	4.25
Japanese	1.3	1.6%	5.0%	3.08
Chinese	14	17.5%	1.3%	0.07

Note: Digital representativeness coefficient = share of website content / population share; a coefficient > 1 indicates overrepresentation, while < 1 indicates underrepresentation. Population data are from Ethnologue (2025); website data are from W3Techs (2025). The collection date was May 6, 2026.

The second imbalance is the further divergence between the share of digital content and the language proportions in the corpora of mainstream large language models. According to official Common Crawl statistics (CC-MAIN-2026-17), among the webpages it crawled, English content accounted for approximately 41.02%, while Chinese content accounted for about 4.86% [21]. The proportion of Chinese in mainstream large language model corpora is likewise mismatched with its population size. At the stage when AI' s "basic worldview" is being formed, narratives and values originating from the English-language internet have already acquired "standard" status through their enormous data weight.

3.2 Training-data landscape of mainstream global large language models

The imbalance in corpus distribution is directly reflected in the training-data composition of mainstream large language models (Table 4). The training data of Western closed-source models are mainly in English, with limited support for multiple languages. Although Western open-source models such as the LLaMA series claim to support multiple languages, English still absolutely dominates their training data. By contrast, Chinese large language models have already achieved breakthroughs in training-data scale and have unique advantages in Chinese-language coverage. DeepSeek-V3 was trained on 14.8 trillion tokens, with corpora mainly in English and Chinese [22]; Tongyi Qianwen 2.5 used 18 trillion tokens, supports more than 29 languages, and performs relatively prominently on Chinese-language tasks.

Table 4 Comparison of training-data scale of mainstream global large language models

Model	Developer	Parameter scale	Training tokens	Language coverage	Open-source status
GPT-4	OpenAI	1.8T (MoE)	13T	Mainly English	Closed source
Claude3.5	Anthropic	Not disclosed	Not disclosed	Mainly English	Closed source
Gemini2.0	Google	Not disclosed	Not disclosed	Multilingual	Closed source
LLaMA3.1	Meta	405B	15T+	12+ languages (mainly English)	Open source (restricted)
DeepSeek-V3	DeepSeek	671B (MoE)	14.8T	Mainly Chinese and English	Open source (MIT)
Qwen2.5	Alibaba	0.5B-72B	18T	29+ languages	Open source (Apache2.0)

Note: Data are from media reports, company technical reports, and public disclosures (2024-2025). The collection date was December 13, 2025.

Through the above comparison, it can be seen that, first, the gap in corpus scale between Western models and Chinese models is narrowing, and the scale of training data for Chinese models has already reached the same order of magnitude as that of leading Western models. Second, the pattern of language coverage is becoming differentiated: Western models use English as the default language, whereas Chinese models have formed a pattern of "Chinese-English bilingualism + multilingual radiation," and have advantages in Chinese-language tasks. Third, China' s leading large models generally adopt highly open open-source strategies, creating advantages for the construction of a global ecosystem.

3.3 Progress in the Construction of Chinese Corpora

While breakthroughs have been achieved at the level of model training, the construction of a high-quality Chinese corpus resource system is showing a trend of multi-level and systematic advancement. At present, a diversified and three-dimensional structure is taking shape, jointly composed of national basic resources, mainstream media resources, social research resources, and industrial application resources (Table 5).

Table 5 Construction of Major Chinese Corpora and Representative Cases

Construction Level	Corpus / Resource System	Scale / Scope	Releasing / Constructing Institution
National basic resources	Chinese Internet Corpus Resource Platform [23]	Integrates 27 authoritative datasets, with a total volume of about 2.7 TB	Guided by the Office of the Central Cyberspace Affairs Commission; China Cyberspace Security Association, National Computer Network Emergency Response Technical Team/Coordination Center of China, etc.
National basic resources	Modern Chinese General Balanced Corpus [24]	Hundreds of millions of Chinese characters	State Language Commission
Mainstream media resources	Mainstream Values Corpus [25]	Contains 30 billion Chinese characters of basic corpus, 300,000 question-answer pairs, and six categories of specialized corpora	National Key Laboratory for Cognitive Communication of Content, People' s Daily
Mainstream media resources	"Shusheng · Wanjuan" Multimodal Corpus [26]	Contains 500 million documents, 22 million image-text items, and more than 1,000 videos	China Media Group, Shanghai Artificial Intelligence Laboratory, etc.

Construction Level	Corpus / Resource System	Scale / Scope	Releasing / Constructing Institution
Social research resources	CLUECorpus2020 [27]	100 GB / about 35 billion characters	Chinese Language Understanding Evaluation (CLUE) organization
Social research resources	Chinese Internet Corpus Series Dataset (CCI) [28]	Total volume about 37 TB, including chain-of-thought (CoT) synthetic data with 420B tokens	Beijing Academy of Artificial Intelligence
Industrial application resources	Private training corpora of leading enterprises (such as those used by DeepSeek, Qwen, and other models)	Trillion-level tokens	DeepSeek, Alibaba, and other technology enterprises

Note: Data are sourced from the National Data Administration, media reports, and public disclosures by organizations; the collection date was December 13, 2025.

At present, at the national level, corpus construction has been elevated to the height of a national language strategy [29]. Mainstream media and professional institutions continue to make efforts in the construction of value-oriented corpora; social research forces focus on the integration of open-source data; and the industrial sector relies on application scenarios to build private corpora and achieve continuous iteration through business feedback. However, the overall operation among different levels still faces the problem of "multiple actors and weak connectivity," and there remains room for improvement in resource integration and efficient utilization.

3.4 Situation Analysis of China' s Corpus Construction

In the competition of corpus-driven intelligent communication, the advantages of China' s corpus construction are mainly reflected in the following four aspects. In terms of data scale, both the training data of frontier models and the construction of basic corpora have reached the trillion-token level, placing

them among the international leaders. In terms of model capabilities, the deep training of domestic models on Chinese corpora gives them a relative advantage when handling Chinese-language and related cultural tasks. In terms of ecosystem strategy, China's leading models generally adopt relatively permissive open-source licenses, which provide favorable conditions for lowering the threshold of use, promoting secondary development by the community, and building a global ecosystem, thereby creating space for the international diffusion of technical standards and value frameworks. In terms of market foundation, the vast

The large base of Internet users and an active application market have provided a solid foundation for the continuous generation of corpora and the iterative development of models.

However, improving China' s capacity for intelligent strategic communication still faces multiple bottlenecks. First, there are problems of resource fragmentation and impeded circulation. Corpus resources held by state institutions, media organizations, platforms, and academic bodies still suffer from the problem of "multiple actors and weak connectivity" ; data systems are not interconnected, constraining resource integration and efficient utilization. Second, there are problems of inconsistent standards and institutional deficiencies. Data formats and annotation specifications are not unified, while systems for property rights and benefit distribution remain vague, raising the cost of integration and hindering the release of the overall strategic value of data elements. Third, there are problems in value alignment and the construction of a cross-cultural communication system. A precise value-expression system for artificial intelligence training remains to be clarified, while a systematic international value-translation system also needs improvement, resulting in a lack of effective narrative carriers when "going global."

The deeper challenge lies in the contradiction facing Chinese-language corpora between "building them well" and "going global." At present, the collection of training data for mainstream large models worldwide still mainly relies on the Western-dominated Internet ecosystem. It is difficult for the high-quality Chinese corpus system built by China to effectively enter the closed-source training processes of leading global artificial intelligence enterprises; this involves multiple constraints such as technical barriers, data sovereignty, and geopolitics. Even for open-source models, their training corpora still take English as the cognitive anchor. Even if Chinese corpora are collected, they may not necessarily acquire sufficient weight in the model-learning process. This is an implicit threshold jointly produced by the current technical system and cultural bias. As a result, China may face the dual risk of being unable to correct mainstream large models' cognitive biases toward China while its own high-quality corpora continue to be marginalized.

The above problems are interconnected and run through every link from foundational resources to strategic output. To resolve these difficulties, systematic planning is required to build a new type of corpus ecosystem that embeds a

clear value orientation, achieves technological interoperability, and can steadily supply high-quality cognitive resources to the world.

4 Practical Pathways for Releasing Intelligent Cognitive Advantages through Corpus Construction

The *Outline of the Fifteenth Five-Year Plan for National Economic and Social Development of the People' s Republic of China* (hereinafter referred to as the *Outline*) states that it is necessary to "uphold the guiding position of Marxism in the ideological sphere, take root in the broad and profound Chinese civilization, follow the trend of information technology development, develop a socialist culture with Chinese characteristics for the new era that has strong ideological leadership, spiritual cohesion, value appeal, and international influence, and solidly advance the building of a culturally strong country." [30] The key to overcoming the predicament of strategic communication in the intelligent era lies in systematically resolving the contradiction between the rich potential resources of Chinese civilization and the insufficiency of effective corpora for intelligent communication. Specifically, it is possible to attempt to construct a collaborative practical pathway of "value—data—governance," thereby systematically transforming profound civilizational resources into high-quality corpora that intelligent systems can recognize, learn from, and recreate.

4.1 Advancing the Construction of a Value-Translation System

At the value level, the goal is to solve the problem of the "computability" of domestic mainstream values in the algorithmic era and to reduce the insufficiency of the "volume" of Chinese discourse in the international arena of public opinion.

First, it is necessary to promote the precision of the value core and build a state-led, machine-readable knowledge graph of mainstream values. The *Outline* states that it is necessary to "construct an identification system for Chinese civilization" [30]. Therefore, interdisciplinary forces should be organized around core contents such as the core socialist values and Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era, transforming them into a structured and computable value knowledge graph, so as to provide a clear value benchmark for corpus annotation and artificial intelligence training.

Second, it is necessary to implement a strategy for translating fine traditional Chinese culture for external communication and to build a cross-cultural mapping network of "value—case—data." Corpus construction needs to carry the function of cross-cultural translation. On the one hand, from the standpoint of the community with a shared future for humanity, it should distill the "[[unclear: phrase beginning with the character '发' is cut off at the bottom of the page]]

"…development," "justice," "peace," "sustainability," and other "value meta-concepts" with a foundation in global consensus, and embed core concepts such as "Chinese-style modernization" and "whole-process people' s democracy" into

the global discourse system for interpretation. On the other hand, it is necessary to establish a deep association mechanism of "value–case–data," concretizing abstract value concepts into factual cases, personal stories, and development data. Through this mapping, when large language models generate content involving China-related issues, they can automatically invoke persuasive factual evidence that has undergone cross-cultural adaptation, thereby replacing the stereotypes commonly found in Western narrative frameworks.

4.2 Promoting the Aggregation and International Interoperability of High-Quality Corpus Resources

At the data level, the goal is to integrate scattered "data islands" through institutional reform, and to resolve the difficulties of insufficient training data and loss of quality control.

First, an institutionalized mechanism led at the national level may be established, with the participation of mainstream media, platform enterprises, and research institutions. This alliance should not merely be a loose cooperative organization, but should become a hub for the allocation of multiple key elements integrating strategic planning, standard-setting, and circulation governance. By formulating unified data standards and interface specifications, high-quality corpora can be integrated and coordinated for use, converging into a national narrative synergy.

Second, the supply of corpus resources should be refined, and an international corpus alliance should be established. The *Outline* states that it is necessary to "deepen exchanges and mutual learning among civilizations, and extensively carry out international cultural and people-to-people exchanges and cooperation" [30]. It is therefore necessary to break through the limitations of the "internal circulation" of corpus collection, proactively integrate overseas Chinese-language media, the research achievements of Sinology in various countries, and materials from "Belt and Road" cooperation projects, and build an internationally interoperable corpus resource repository. By proactively providing global AI models with undistorted raw materials from a Chinese perspective, cognitive and value biases in strategic communication can be corrected at the source. At the same time, corpora entering the repository should undergo strict labeling and quality grading across multiple dimensions, including political orientation, authenticity, and cultural appropriateness, forming a complement to existing domestic corpus systems and realizing both the "internal circulation" and "external circulation" of a high-quality corpus system of Chinese civilization.

4.3 Promoting Innovation in Proactive and Collaborative Corpus-Governance Models

At the governance level, the goal is to grasp the power to define discourse in the intelligent era through forward-looking, source-oriented rule guidance and agile governance.

First, governance should be made anticipatory, source-oriented, and agile. It must be recognized that corpora are the "new gatekeepers" of the intelligent era, filtering knowledge and defining cognitive benchmarks. The focus of governance should shift from end-stage content screening to compliance review at the data source and value alignment of algorithmic models, increasing the weight of high-quality corpora in training and enhancing transparency disclosure. A full-process audit mechanism for AI-generated content should be established, embedding governance norms into every link of training and inference, so as to move from passive response to active guidance.

Second, efforts should be made to explore and build a co-governance mechanism led by the government and involving diverse social actors, to supervise corpus construction and AI applications, prevent excessive concentration of technological power, and, on this basis, actively participate in the formulation of global artificial-intelligence governance rules, avoiding a disadvantaged position in rule-making and gaining discourse power in governance.

5 Conclusion

In the intelligent era, the competitive field of strategic communication has shifted from "surface-level narratives" to "underlying corpora." This study defines the core category of "intelligent strategic communication capacity" and demonstrates the internal logic by which corpus construction transforms from a mere data resource into a strategic cognitive infrastructure. However, the study still has certain limitations: first, as an exploratory study in theoretical construction, it focuses on the conceptual framework, operating mechanisms, and macro-level pathways of "intelligent strategic communication capacity," and has not yet conducted quantitative measurement or dynamic analysis of the strength or weakness of this capacity.

comparison. Second, against the backdrop of rapid iteration in intelligent technologies, the new opportunities and challenges that may arise from the paradigmatic evolution of corpus construction and model training warrant further examination in terms of their dynamism. Looking ahead, it is necessary to further construct a multidimensional, observable evaluation system covering the three-layer framework of "generative foundations—cognitive coordination—generative efficacy," and to combine longitudinal tracking or cross-national comparative cases to empirically test the effectiveness of different pathways, thereby further advancing this field from theoretical construction toward measurement-based assessment and practical optimization.

【References】

[1] Hu Yiqing, Hu Xinyue. ChatGPT as language infrastructure [J]. Journalism Review, 2023(6): 21–27.

[2] Shi Anbin, Tong Tong. From international communication to strategic com-

munication: contextual adaptation and path transformation in the new era [J]. News and Writing, 2021(10): 14-22.

[3] Weller T, Lock I. Automated communication' s impact on strategic communication: implications from a systematic review [J]. International Journal of Strategic Communication, 2025, 19(1): 13-34.

[4] Hallahan K, Holtzhausen D, Van Ruler B, et al. Defining strategic communication [J]. International Journal of Strategic Communication, 2007, 1(1): 3-35.

[5] Holtzhausen D R, Zerfass A. Routledge handbook of strategic communication [M]. New York: Routledge, 2014.

[6] Miskimmon A, O' loughlin B, Roselle L. Strategic narratives: communication power and the new world order [M]. Routledge, 2014: 1-4.

[7] Jiang Fei, Zhang Nan. Three waves of China' s international communication (1978-2019) [J]. Global Media Journal, 2019, 6(2): 39-58.

[8] Noble S U. Algorithms of oppression: how search engines reinforce racism [M]. New York: New York University Press, 2018.

[9] Thussu D K. International communication: continuity and change [M]. 2nd ed. London: Hodder Education, 2006.

[10] Zhang Yaoqiang, Pang Min. China' s international communication in the new era: new starting points, new logic, and new paths [J]. Modern Communication (Journal of Communication University of China), 2021, 43(7): 40-49.

[11] Bender E M, Gebru T, McMillan-Major A, et al. On the dangers of stochastic parrots: can language models be too big? [C] //Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. 2021: 610-623.

[12] Gallegos I O, Rossi R A, Barrow J, et al. Bias and fairness in large language models: a survey [J]. Computational Linguistics, 2024, 50(3): 1097-1179.

[13] Perelkiewicz M, Poświata R. A review of the challenges with massive web-mined corpora used in large language models pre-training [C] //International conference on artificial intelligence and soft computing. Cham: Springer Nature Switzerland, 2024: 153-163.

[14] Xu Y, Hu L, Zhao J, et al. A survey on multilingual large language models: corpora, alignment, and bias [J]. Frontiers of Computer Science, 2025, 19(11): 191136.

[15] Tao Y, Viberg O, Baker R S, et al. Cultural bias and cultural alignment of large language models [J]. PNAS Nexus, 2024, 3(9): pgae346.

[16] Wang Z, Wu Z, Zhang Y, et al. Bias amplification: large language models as increasingly biased media [C] //Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-

Pacific Chapter of the Association for Computational Linguistics. 2025: 115-132.

[17] Zhu L, Mou W, Lai Y, et al. Language and cultural bias in AI: comparing the performance of large language models developed in different countries on traditional Chinese medicine highlights the need for localized models [J].

Han Bo. Research on the Value and Pathways of Empowering National Intelligent Strategic Communication Capacity through High-Quality Corpora [J]. Journal of Documentation and Data, 2026, 8(2): 016-028.

Journal of Translational Medicine, 2024, 22(1): 319.

[18] Rystrøm J H, Kirk H R, Hale S A. Multilingual!= multicultural: evaluating gaps between multilingual capabilities and cultural alignment in LLMs [C] //Proceedings of Interdisciplinary Workshop on Observations of Misunderstood, Misguided and Malicious Use of Language Models. 2025: 74-85.

[19] Eberhard D M, Simons G F, Fennig C D. Ethnologue: languages of the world [M/OL]. 28th ed. Dallas, Texas: SIL International, 2025 [2025-12-13]. https://www.ethnologue.com/.

[20] W3Techs. Usage statistics of content languages for websites [EB/OL]. (2025-12-12) [2025-12-13]. https://w3techs.com/technologies/overview/content_language.

[21] Common Crawl. Statistics of Common Crawl monthly archives: distribution of languages [EB/OL]. [2026-05-06]. https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.

[22] DeepSeek-AI. DeepSeek-V3 technical report [EB/OL]. (2024) [2025-12-13]. https://arxiv.org/abs/2412.19437.

[23] Xinhua News Agency. Chinese internet corpus resource platform released [EB/OL]. (2025-01-10) [2025-12-13]. https://www.news.cn/tech/20250110/26dca8ee795d4625a647bb6d92db15

[24] China Language and Writing Network. Modern Chinese corpus of the State Language Commission [EB/OL]. [2025-12-13]. http://www.china-language.edu.cn/#/languageResources/corpus.

[25] National Data Administration. Typical cases of high-quality datasets | Mainstream value corpus [EB/OL]. (2025-10-10) [2025-12-13]. https://www.nda.gov.cn/sjj/ywpd/szkjyjcss/10

[26] Shanghai Artificial Intelligence Laboratory. The Large-Model Corpus Data Alliance openly releases the high-quality multimodal corpus "Shusheng Wanjuan"[EB/OL]. (2023-08-14) [2025-12-13]. https://www.shlab.org.cn/news/5443473.

[27] Xu L, Zhang X, Dong Q. CLUECorpus2020: a large-scale Chinese corpus for pre-training language model [EB/OL]. (2020-03-03) [2026-05-03]. https://arxiv.org/abs/2003.01355.

[28] National Data Administration. Typical cases of high-quality datasets | Chinese Internet Corpus series datasets (CCI) [EB/OL]. (2025-10-10) [2025-12-13]. https://www.nda.gov.cn/sjj/ywpd/szkjyjcss/1010/20251010184420955142789_pc.html.

[29] Ministry of Education, State Language Commission, Cyberspace Administration of China. Opinions on strengthening digital Chinese construction and promoting the development of language and writing informatization [EB/OL]. (2025-01-08) [2025-12-19]. https://www.gov.cn/zhengce/zhengceku/202503/content_7016543.htm.

[30] *Outline of the Fifteenth Five-Year Plan for National Economic and Social Development of the People's Republic of China* [M]. Beijing: People's Publishing House, 2026: 91, 94, 99.

High-Quality Corpus for National Intelligent Strategic Communication: Value and Paths

Han Bo

(Institute of Journalism and Communication, Chinese Academy of Social Sciences,
Beijing 100720, China)

Abstract: [**Purpose/Significance**] The competitive focus of international strategic communication is shifting from the content level to the foundational level of training corpora that shape AI cognition. From the perspective of high-quality corpus construction, this study explores its mechanism for empowering national intelligent strategic communication capacity and the practical paths for releasing intelligent cognitive advantages. [**Method/Process**] This paper proposes the concept of "intelligent strategic communication capacity," which refers to the capability of actors to shape the internal knowledge framework and value logic of artificial intelligence models through the systematic construction of cognitive infrastructure such as corpora. Based on open data on large language model corpora, it comparatively analyzes the linguistic distribution of global corpus resources and the compositional characteristics of the training data of mainstream models, evaluates China's cognitive status, and proposes practical paths in conjunction with policy text analysis. [**Result/Conclusion**] As a core resource for intelligent strategic communication, corpora exert a double-edged-sword effect. On the one hand, the dual imbalance of global corpus resources is triggering systematic intelligent cognitive biases that are unfavorable to non-Western countries. On the other hand, strategic corpus construction can transform such disadvantages into strategic intelligent cognitive advantages. At present, China's corpus construction still faces structural dilemmas, including data silos, the lack of unified standards, translation distortion, and difficulty in incorporating high-quality corpora into the training process of mainstream global models. This study proposes a tripartite collaborative governance path integrating values, data, and governance, so as to provide references for smoothing the transformation logic from civilizational values to corpus resources and further to strategic communication effectiveness.

Keywords: Intelligent strategic communication capacity; Corpus; Generative AI; International strategic communication

---

(Responsible editor: Kong Qingqing)

Note: Figure translations are in progress. See original paper for figures.