Computers and Artificial Intelligence; Vol.3, No.4; 2026
E-ISSN: 3006-3728 P-ISSN: 3006-371X
DOI: https://doi.org/10.70267/cai.v3n4.193206
Published by: Zeus Press

A Review of Retrieval-Augmented Generation: Architectural Evolution, Key Breakthroughs, and Vertical-Domain Practices

Zihe Zhu and Bin Chen*

College of Hydraulic Science and Engineering, Northeast Agricultural University, Harbin, 150030, China

**Corresponding author: Bin Chen.*

Abstract

Large language models (LLMs) rely on massive pretraining based on the Transformer architecture and show significant emergent effects in general tasks. However, the nature of their statically parameterized knowledge causes vertical-domain applications to face three core bottlenecks: knowledge hallucination, delayed timeliness, and loss in long contexts. Retrieval-Augmented Generation (RAG) realizes the decoupling of computational logic and factual memory through the paradigm of "retrieve first and generate afterward". It can dynamically inject domain knowledge without fine-tuning and has become a mainstream technical path for the implementation of trustworthy artificial intelligence (AI). This study systematically reviews the evolution of RAG technology from 2024 to 2026, from Naive RAG to Advanced RAG, Modular RAG, and graph-enhanced RAG. It analyzes in depth the breakthrough progress of four key technologies, namely hybrid retrieval, multimodal document parsing, generation control, and lightweight deployment. Combined with typical application cases in complex equipment manufacturing, energy and power, public services, and other fields, it explains the engineering value of RAG. Finally, it discusses the evaluation system, existing limitations, and future trends, providing a reference for academic research and practical implementation of RAG technology.

Keywords

retrieval-augmented generation, large language model, knowledge graph, multimodal fusion, vertical-domain application

1. Introduction

In recent years, AI technologies represented by large language models (LLMs) have made breakthrough progress. Their performance in natural language understanding, generation, and complex reasoning tasks has profoundly changed the mode of human-computer interaction. However, because LLMs rely on static parameters to store knowledge, they have inherent defects when facing vertical-domain applications. First, there is the problem of knowledge hallucination, in which models easily generate content inconsistent with facts. Second, timeliness lags behind, and the latest domain knowledge cannot be incorporated in real time. Third, long-context loss occurs, and key information may be omitted or semantic shifts may appear when ultra-long documents are processed. These defects seriously restrict the large-scale application of LLMs in scenarios with extremely high requirements for accuracy and reliability, such as medicine, industry, and finance. Problems such as different answers to the same question and unsupported expansion in enterprise institutional

© 2026 The Author(s). Authors retain copyright in their articles.

scenarios have been empirically summarized [1], and RAG has also been used as a lightweight anti-hallucination solution in cybersecurity and vulnerability awareness scenarios [2, 3].

From 2024 to 2026, RAG rapidly moved from general question answering to vertical domains, covering key scenarios such as education, energy and power, public safety, and industrial operation and maintenance, and forming many reusable engineering solutions [2, 4, 5]. To solve the above problems, the Meta AI team formally proposed Retrieval-Augmented Generation (RAG) in 2020 [6]. Its core idea is to combine the retrieval results of an external knowledge base with the generation capability of an LLM, so that the model can preferentially refer to authoritative external information when generating answers, thereby decoupling "computational logic" from "factual memory". Compared with traditional fine-tuning methods, RAG can dynamically inject new knowledge without updating model parameters, greatly reducing the cost of domain adaptation while improving the interpretability and traceability of answers. Since 2024, RAG technology has entered a period of rapid evolution. It has gradually developed from early Naive RAG into advanced forms integrating knowledge graphs, multimodal parsing, localized deployment, and agent architectures, and has formed many reusable engineering solutions in fields such as complex equipment development, energy and power dispatching, and medical and health consultation [7, 8, 9]. This study systematically reviews the current research status of RAG technology from four dimensions: architectural evolution, key technologies, domain applications, and evaluation systems, and it also looks ahead to future development directions.

2. Basic Architecture and Evolutionary Lineage of RAG

2.1 End-to-End Full-Link Basic Pipeline

A typical RAG system builds a full-process closed loop from knowledge processing to answer generation. Its core links include four stages: knowledge embedding, retrieval, augmentation, and generation [10, 11, 12]. In the knowledge embedding stage, unstructured or semi-structured external knowledge is segmented, converted into high-dimensional vectors through an embedding model, and stored in a vector database [10, 11, 12]. The quality of this link directly determines the recall rate of subsequent retrieval, especially because difficult extraction of complex elements such as tables, formulas, and charts must be solved [11, 13]. The core formula of RAG marginalizes the retrieved documents and eliminates the latent variable of "which document is used". It does not select the "only correct document". Instead, it assigns a weight to each candidate document, integrates all possible sources in a weighted manner, and then generates. In essence, it is a probabilistic expression of "retrieve first and generate afterward". This is shown in Formula 1.

$$P(y|x) = \sum_{x \in c} P(z|x)P(y|x,z) \quad (1)$$

Where

x: input query

z: document retrieved from the external corpus (paragraph/chunk)

y: generated answer

$P(z|x)$: document relevance probability given by the retriever

$P(y|x,z)$: generation probability under the condition of "query + document"

In the retrieval stage, the user query is converted into a vector through the same embedding model, and Top-K relevant text chunks are recalled from the vector database based on cosine similarity calculation [14, 15]. Basic retrieval strategies mainly rely on vector similarity, but they are easily affected by semantic ambiguity and therefore are often combined with keyword matching methods [16, 17]. In the augmentation stage, prompt engineering is used to concatenate retrieval results with the original query, and constrained prompts such as "answer the question based on the following authoritative information" are designed to guide the LLM to preferentially refer to external knowledge rather than parameterized memory [10, 18]. This link needs to focus on solving the constraints of the context window and the filtering of noisy information [19, 20]. In the generation stage, the augmented prompt is input into the LLM. The model combines the retrieved authoritative knowledge with its own language capability to generate the final answer. Some systems also attach provenance markers, such as cited document IDs and paragraph locations, to improve interpretability

© 2026 The Author(s). Authors retain copyright in their articles.

[7, 21, 22]. The advantages of this basic pipeline are strengthened in enterprise institutional governance scenarios. Through a closed-loop mechanism of "refusal-answer, tracing-back, and escalation", semantic consistency, factual consistency, and auditability are unified [1].

2.2 Architectural Evolution from Flat Retrieval to Multi-Engine Collaboration

RAG technology is widely applied in many fields, as shown in Table 1.

Table 1: Applications in Different Fields

Field	Representative system	Knowledge source	Retrieval strategy	Generation model	Core indicator/effect
Enterprise governance	Consistency iceberg model [1]	Institution texts/processes/versions	Semantic retrieval + refusal tracing	LLM + RAG	Consistency in semantics, facts, and auditability
Cybersecurity	Vul-RAG class [2, 3]	CVE/threat intelligence	Cross-domain vector + metadata	Local LLM	Improved semantic quality of vulnerability risk
Education	Classroom assistant [23, 24]	Textbooks/question banks/ISTQB	BM25 + vector + Cross-Enc.	DeepSeek/Qwen	Hallucination rate <5%; accuracy 84%
Course KG-QA	LightRAG fusion [25, 26]	Textbook KG + vector	KG path + vector hybrid	GLM-4.5	Retrieval 94.5%; QA 93.2%
Industrial O&M	DS^2O-Graph-RAG [27, 28]	Equipment manuals + graph	Structural reranking + hierarchy	Single LLM call	Fuzzy-input accuracy >90%
Government/grass roots	ChatGovt [29, 30]	Policies/work orders/cases	Historical summary + reranking + reflection	Qwen2.5-14B	Recall +55.4%; reporting integration
Infrastructure	Railway four-electricity [5]/BIM [31]	Regulations/IFC models	NL2Cypher + flattened attributes	Qwen3 + LoRA	Top-5 recall 91%; mapping <812 ms
Medicine/safety	Chinese patent medicine recommendation]/hazardous chemicals [4]	Pharmacopoeia/accident cases	Dynamic cloud knowledge-base update	Tongyi Qianwen-Max	Micro-F1 +30.16%

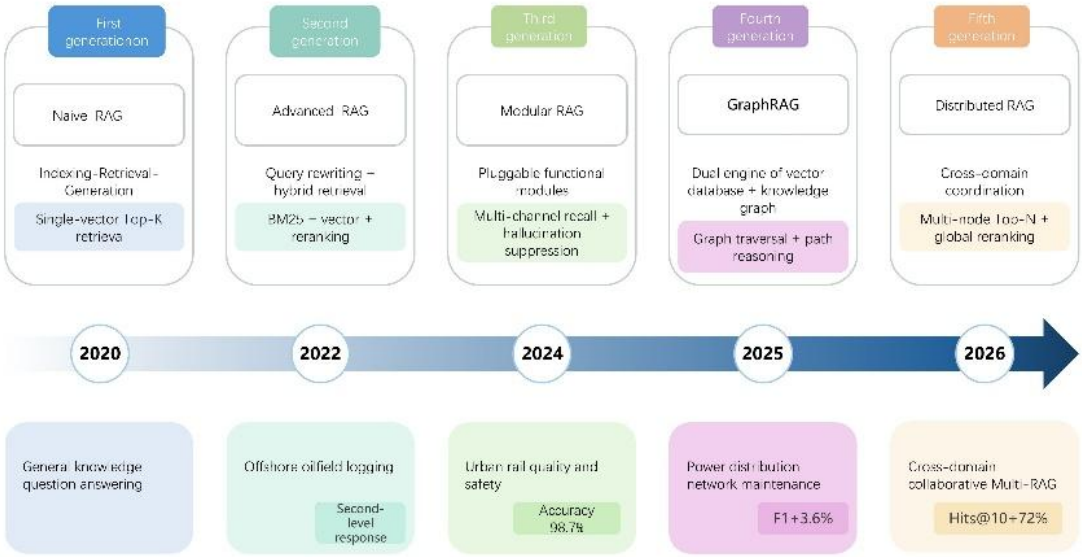
With the increasing complexity of application scenarios, the RAG architecture has experienced three important generations of evolution and has gradually formed a modular and customizable engineering system. Early Naive RAG follows the linear pipeline of "indexing, retrieval, and generation". It only supports vector retrieval of unstructured text and is suitable for simple question-answering scenarios such as general knowledge queries [14, 32]. However, in vertical domains, it faces problems such as shallow semantic understanding and insufficient logical reasoning capability. For example, the terminology recognition rate in ship industry applications is lower than 70% [33]. Classroom intelligent assistants in the education field also expose the problem that single vector retrieval provides insufficient support for professional terms and formulas, which has promoted the popularization of hybrid retrieval and reranking mechanisms in Advanced RAG [23, 34].

To make up for these defects, Advanced RAG adds preprocessing and postprocessing optimization modules on the basis of Naive RAG. In the preprocessing stage, a query rewriting mechanism is introduced to expand

© 2026 The Author(s). Authors retain copyright in their articles.

semantics and solve the problem of vague user query expressions [16,35]. In the retrieval stage, a hybrid retrieval strategy is adopted to integrate Best Matching 25 (BM25) sparse retrieval, that is, keyword matching, with dense vector retrieval, that is, semantic matching [8, 17]. In the postprocessing stage, a Cross-Encoder reranking model is introduced to improve the relevance of retrieval results [8, 15]. After an offshore oilfield logging system applied this architecture, it realized second-level responses and improved engineers' decision-making efficiency by 40% [8]. Modular RAG further decomposes the system into pluggable functional modules and supports flexible combinations according to scenario requirements. Its core modules include a query understanding module, including intent recognition and entity linking; a multi-channel recall module, including vector retrieval, graph retrieval, and keyword retrieval; a fusion ranking module, including rule fusion and learning-to-rank; and a generation control module, including hallucination suppression and evidence verification [20]. In grass-roots governance scenarios, the grid question-answering assistant adopts this type of modular design, encapsulating document segmentation, vector retrieval, reranking, and structured reporting into a reusable link to realize the integration of "caliber query + element extraction + reporting" [30]. This modular design can adapt to complex industrial scenarios. For example, in an urban rail quality and safety management system, through the multi-module collaboration of "text retrieval + image-text association + expert experience matching", the accuracy of text retrieval can reach 98.7% [36]. To address the defect that traditional RAG has difficulty processing structured relational reasoning, graph-enhanced RAG, including GraphRAG and KRAG, integrates knowledge graphs with vector databases to form a "dual-engine" architecture. The vector database is responsible for semantic similarity matching, and the graph database is responsible for entity-relation reasoning, supporting multi-hop queries, such as "equipment fault → related maintenance record → corresponding solution" [7, 9, 37]. The distribution network maintenance framework uses graph path reasoning to improve the F1 value of maintenance decision-making by 3.6% compared with traditional RAG [38]. The aquatic nutrition recommendation system uses Neo4j graph paths to forcibly prohibit diets for patients with gout, effectively solving the problem of "missing logical constraints" in vector retrieval [37]. In equipment operation and maintenance and building information modeling (BIM) engineering audit scenarios, structural semantic collaborative optimization (DS^{2}O-Graph-RAG) and Natural Language to Cypher (NL2Cypher) mapping further solve the retrieval drift problem under fuzzy input and increase the accuracy under entity redundancy and missing-entity scenarios to more than 90% [27, 31]. The iteration of RAG technology is shown in Figure 1. Different RAG categories are shown in Table 2.

Figure 1: Evolution map of the RAG architecture



© 2026 The Author(s). Authors retain copyright in their articles.

Table 2: RAG Categories

Gen.	Retrieval	Knowledge organization	Generation control	Defect fix
Naive	Single-vector Top-K	Flat text chunks	Prompt concatenation	Semantic ambiguity; no provenance
Adv.	BM25 + vector + reranking	Metadata chunks	Confidence threshold	Term variation
Modular	Multi-channel recall (graph/keyword/vector)	Pluggable modules	Hallucination/evidence check	Module consistency
Graph	Graph traversal + vector fusion	Neo4j triples	Path-constrained generation	Fuzzy entity alignment
Distributed	Multi-node Top-N + global reranking	Cross-domain independent indexing	No raw data aggregation	Collaboration under privacy compliance [39]

3. Key Technological Breakthroughs of RAG

3.1 Hybrid Retrieval: Balancing Exact Matching and Semantic Generalization

A single retrieval method has natural limitations. Sparse retrieval methods such as BM25 are good at exact matching of professional terms, but they have difficulty understanding semantic associations. Dense vector retrieval is good at capturing semantic associations, but it is sensitive to keyword spelling errors and terminology variants [16, 17]. To break this bottleneck, industry has explored a variety of hybrid retrieval strategies and realized complementarity. The railway four-electricity knowledge-enhanced retrieval system adopts a "structural tag + vector recall" mode, adding tags such as "equipment type" and "operation link" to technical documents and significantly improving retrieval precision [5]. A similar model is also applied in the offshore oilfield logging system OffshoreRAG. The system adds structural tags such as "equipment type", "operation link", and "fault code" to documents such as logging reports and well-test designs. During retrieval, it first narrows the scope through tag filtering and then performs vector similarity matching, which significantly improves the retrieval precision of unstructured documents [8]. The "BM25 + vector + reranking" mode has been verified in the visual symbolic reasoning framework RAG-SymbolicDet. Experiments show that hybrid retrieval reaches an AUROC of 0.8640 in the helmet compliance detection task, outperforming single sparse retrieval (0.8215) and single dense retrieval (0.8423). Fine-grained reranking of hybrid retrieval results through a Cross-Encoder can further improve the ranking priority of key evidence [17]. On this basis, an intelligent question-answering system for Python programming introduces an adaptive weight allocation algorithm to dynamically adjust the fusion ratio of BM25 and vector retrieval, increasing BLEU, ROUGE-L, and accuracy by 10.3%, 12.10%, and 11.20%, respectively [34]. For policy consultation scenarios where structured fields and unstructured descriptions coexist, the "Text2SQL + vector" mode provides a precise solution. It uses a large model to convert user queries into SQL statements, accurately extracts structured policies from a relational database, and also vectorizes the query to retrieve unstructured supplementary explanations. The system gives priority to SQL results and falls back to vector results when the SQL result is empty, avoiding the problem that traditional RAG returns irrelevant policies through "similarity retrieval" [15]. In addition, the domain dictionary enhancement mode has also been successfully applied in ship industry RAG systems. By constructing an enhanced dictionary containing professional terms such as "load line" and "capacity coefficient", terminology standardization is performed before vector retrieval, for example unifying "ship shell plate" as "outer plate", increasing the terminology recognition rate to 93.7% [33].

3.2 Multimodal Document Parsing: Solving the Problem of Heterogeneous Data Integration

Vertical-domain knowledge often exists in multimodal forms, such as CAD drawings, formulas, and tables in engineering, imaging reports and inspection curves in medicine, and soil sample images and meteorological data tables in agriculture [11, 18, 40]. Traditional parsing tools have difficulty processing complex layouts, resulting in low-quality knowledge-base construction. To break this bottleneck, recent studies have turned to more adaptive hybrid parsing frameworks. Taking the general document parsing framework UniDoc-Parser as an example, it integrates multiple parsing algorithms and can adaptively call the optimal path according to file type. For PDF, it uses a layout parsing model to identify titles, paragraphs, tables, and image regions. For DOCX, it can directly extract structured XML. For XLSX, it converts the file into CSV and attaches table

© 2026 The Author(s). Authors retain copyright in their articles.

structure descriptions. For scanned documents, it uses optical character recognition (OCR) and layout restoration [27]. Experiments show that this framework is superior to mainstream open-source tools in table structure recognition (TEDS = 69.3%) and named entity extraction (NED = 63.8%). In power design, a scenario with extremely high precision requirements, the strategy of "layout parsing + LLM recursive chunking" has remarkable effects. It first uses the DocLayout model to identify 23 document layout categories, such as titles, body text, tables, and formulas, and filters noise such as headers and footers. Then it uses Sentence-BERT to calculate the similarity between text chunks and title chunks and correct titles that have been misclassified. Finally, it performs recursive segmentation based on the document hierarchy to ensure semantic integrity within chunks [13]. This method improves downstream retrieval accuracy by 22% and the F1 value of the question-answering system by 18%. In an intelligent ancient-book ontology system, the ONT2Vec hybrid embedding model for ancient-book ontology structure design jointly represents textual semantic embeddings and graph structure embeddings, improving the accuracy of ontology question answering and alignment tasks. To meet professional modal fusion requirements, the ship pipeline design system simultaneously processes text specifications, two-dimensional drawings, and three-dimensional models. It uses Jieba segmentation plus a ship professional dictionary for text, OCR to identify dimension annotations and valve symbols in drawings, and extracts spatial coordinates and connection relationships from models. Finally, it realizes semantic alignment of "text-image-model" through the Contrastive Language-Image Pretraining (CLIP) model [33, 40]. The offshore oilfield logging system specially optimizes feature extraction of logging curves, converting curve forms, such as peaks and slopes, into semantic descriptions, such as "signs of lost circulation", and incorporating them into the vector retrieval scope [8].

3.3 Generation Control: Suppressing Hallucination and Ensuring Trustworthiness

Although RAG can alleviate hallucination, it may still generate incorrect information when retrieval quality is low and context noise is large. For this reason, studies have proposed a variety of generation control mechanisms. In the soil acidification diagnosis system "Sinong + RAG", a strict evidence verification mechanism is embedded. Before generating an answer, the system scores the confidence of retrieval results. When relevant evidence is insufficient, it refuses to generate and prompts "insufficient information; please supplement soil pH data", while triggering a manual review process [18], thereby cutting off the spread of incorrect information at the source. The "generation-evaluation-optimization" closed-loop mechanism (ED-RAG) proposed for coal mine electromechanical equipment fault diagnosis introduces an independent LLM reviewer to score the initial answer from multiple dimensions and guide iterative optimization. Finally, the average win rate of the optimized answer reaches 63.7% [28]. The misinformation detection system requires generated answers to cite evidence from at least two independent sources; otherwise, the answer is marked as "to be verified" [41, 42]. The confidence-driven dynamic expansion mechanism is adopted by the DAR-RAG framework. It introduces multidimensional confidence evaluation, including retrieval relevance, evidence consistency, and logical coherence. When confidence falls below the threshold, external knowledge sources, such as real-time databases and expert experience bases, are dynamically called to expand information coverage and avoid erroneous generation under low-quality retrieval [20]. In more complex decision-making scenarios, self-evaluation feedback is especially important. The community security risk disposal system (HG-RAG) embeds a self-checking logic for "closure of the disposal process". After generating recommendations, it automatically checks whether the full chain of "on-site disposal → system reporting → tracking feedback" is covered. If not, it retrieves and optimizes again. This mechanism improves the decision-value score by 12% [43]. The government hotline system ChatGovt introduces historical dialogue summaries and self-reflection. While ensuring traceability, it greatly improves answer recall by 55.4% compared with fine-tuned GLM-4-9b-chat [29]. For the risk of data tampering after retrieval, the trusted smart patent data system uses dual-path verification of T2S (text to graph) and S2T (graph to text). Retrieval results are verified for consistency through both the vector database and the graph database, and tampered content is filtered because it cannot match graph entity relationships [44].

3.4 Localized and Lightweight Deployment: Adapting to Data Privacy and Resource Constraints

In fields with strict data privacy requirements, such as military industry, power, and medicine, together with the limited computing power of some edge scenarios, RAG is developing toward localization and lightweight deployment. In the adaptation of domestic models, a university library reference consultation

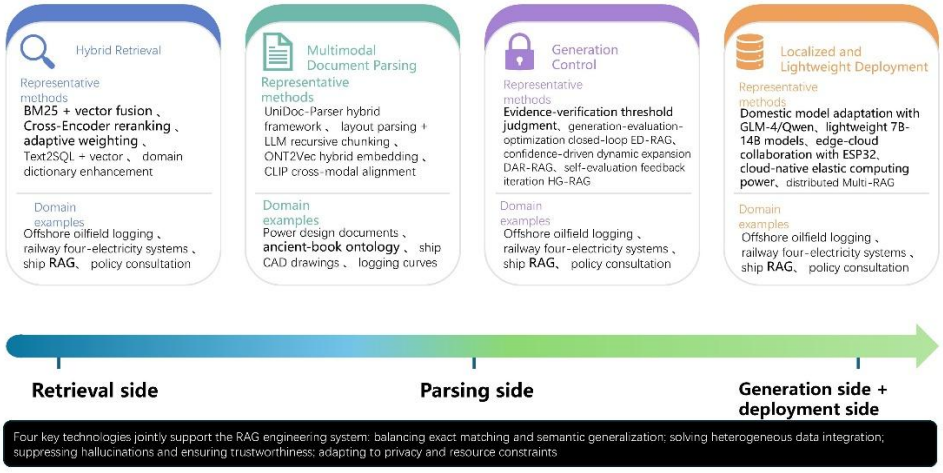
© 2026 The Author(s). Authors retain copyright in their articles.

system deploys the Zhipu GLM-4 model based on the domestic Baidu PaddlePaddle AI Studio platform. Tests on 906 real queries show an accuracy of 0.82 and a response time of 12.7 seconds, with overall performance superior to cloud models such as ChatGPT [22]. The ship industry RAG system adopts a fully domestic technology stack, from chips to models, and its inference efficiency is 3.2 times higher than that of cloud-based solutions [33]. The combination of lightweight models and RAG has been verified in a lung cancer specialty question-answering system. A comparison of 11 domestic models from 7B to 32B finds that the introduction of RAG significantly improves the accuracy of all models. Among them, DeepSeek-7B increases from 32.50% to 60.00% ($P < 0.001$), while Qwen3-14B achieves the best balance among accuracy (81.5%), response latency (11.2 seconds), and computing power requirements, making it suitable for local deployment in hospitals [45]. Facing complex industrial sites, an edge-cloud collaborative architecture is applied in an electric spark leak detector. The device integrates an ESP32 microcontroller and locally stores an anti-corrosion coating detection knowledge base, including standards, specifications, calculation formulas, and fault cases. Through RAG, it realizes voice interaction and automatic parameter calculation. Complex problems, such as detection of new coating materials, are collaboratively inferred by a large model on a local server, balancing real-time performance and intelligence [46]. The Chinese patent medicine intelligent recommendation system adopts a cloud-native architecture. Through the elastic computing power of the Alibaba Cloud Bailian platform and the Tongyi Qianwen-Max-Latest model, it realizes dynamic updating of the traditional Chinese medicine knowledge base while significantly reducing local computing costs [47]. The Multi-RAG framework for cross-domain data goes further. Through multi-node independent indexing and global reranking, it achieves Hits@10 of 0.7659 without moving raw data out of the domain, a 72% improvement over a single node [39]. The comparison of different methods on unified data is shown in Table 3. The classification of RAG core technologies is shown in Figure 2.

Table 3: Comparison of Different Methods on Unified Data

Task/dataset	Comparison methods	Best method
Visual symbolic reasoning (helmet detection AUROC)	Sparse retrieval 0.8215 / dense retrieval 0.8423 / hybrid retrieval 0.8640 / + reranking 0.8786	RAG-SymbolicDet + reranking
Python question answering (relative BLEU improvement)	Baseline / BM25 + vector / + adaptive weight fusion +10.3%	Adaptive weight fusion
Distribution network maintenance QA (F1)	No RAG 0.704 / traditional RAG 0.704 / KRAG-LLM-DM 0.720	Graph-enhanced KRAG
Lung cancer specialty QA (7B accuracy)	DeepSeek-7B 0.325 / + RAG 0.600 / Qwen3-14B + RAG 0.815	Local lightweight Qwen3-14B + RAG
Urban rail corpus governance (retrieval accuracy)	Traditional RAG 0.920 / multi-module collaboration 0.987	Multi-module collaborative RAG
Cross-domain collaboration (Hits@10)	Single node 0.4453 / Multi-RAG 0.7659	Multi-RAG (+72% over single node)

Figure 2: Classification of RAG Key Technologies



© 2026 The Author(s). Authors retain copyright in their articles.

4. Typical Vertical-Domain Application Map of RAG

At present, RAG technology has been implemented in six major directions: enterprise governance, cybersecurity, education, industrial operation and maintenance, public services, and cultural-tourism planning. In enterprise institutional scenarios, the consistency iceberg model is used to reduce communication costs [1]. In cybersecurity, lightweight RAG is used to realize vulnerability risk awareness and threat analysis [2, 3]. Education scenarios cover the complete chain from classroom assistants to course knowledge graph question answering [23, 24, 25, 26]. Industrial operation and maintenance and engineering infrastructure scenarios use Graph-RAG and NL2Cypher technologies to solve equipment diagnosis and BIM audit problems [5, 27, 31]. Public service scenarios include government hotlines, grass-roots grids, and military equipment question answering [29, 30, 48]. Cultural-tourism and medicine scenarios respectively realize personalized travel route generation and intelligent Chinese patent medicine recommendation [47, 49], integrating the technology into daily applications.

4.1 Complex Equipment Development

Complex equipment development, from requirement demonstration, architecture design, and physical simulation to testing and verification, faces pain points such as tool heterogeneity, data silos, and difficulty in traceability [7]. The RAG-based model-based systems engineering (MBSE) data foundation adopts a four-layer architecture: computing-power foundation layer, data preprocessing layer, dual-engine knowledge storage layer (vector database + graph database), and application hub layer [7]. Key technologies include multimodal semantic alignment of heterogeneous engineering entries, business-process-driven vector-graph hybrid mapping, which constructs a "requirement-function-logic-physical" traceability matrix, and lightweight RAG multi-hop reasoning [7]. An application in a military equipment development project shows that this architecture shortens design-change response time by 40% and improves traceability efficiency by 65%. The ship pipeline design system integrates knowledge graphs and vector retrieval, associates international specifications, design parameters, and engineering cases in multiple dimensions, and supports the closed loop of "query-design-verification", reducing the design cycle by 25% compared with before [40]. The ship professional vocabulary-enhanced RAG system constructs a multimodal document parsing pipeline and optimizes terminology recognition for heterogeneous data such as CAD drawings and maintenance manuals. In a 1.2 GB historical data test, retrieval accuracy reaches 89.2%, and shipowner question-answering response time reaches the millisecond level [33]. The finite element modeling assistance system constructs a RAG knowledge base based on ABAQUS software operation logs. The accuracy of generated modeling commands is 52% higher than that of direct LLM generation, supporting rapid parametric modeling of complex hull structures [50].

4.2 Energy and Power

Power systems have extremely high requirements for knowledge accuracy and timeliness, and RAG has become a core technology for grid intelligence. The distribution network maintenance auxiliary decision-making framework KRAG-LLM-DM innovates a knowledge graph construction method of "large-model guidance + few-shot learning", automatically extracting 18,280 maintenance entities and 14,002 relationships. In the maintenance question-answering task, its F1 value reaches 0.72, 1.60% higher than the non-RAG baseline [38]. The power quality retrieval engine EPath-RAG introduces professional term rewriting, two-layer keyword extraction, and knowledge graph path reasoning, shortening query response time to 0.8 seconds [9]. The intelligent regulation method for power system frequency combines RAG with low-frequency load-shedding rules and stores optimization rules through a vector database. Simulation shows that, compared with traditional methods, frequency recovery time is shortened by 35% and steady-state frequency deviation is reduced by 0.12 Hz [51]. In the railway four-electricity field, the dual-route collaborative framework of "RAG + low-rank adaptation (LoRA) domain fine-tuning" shortens frontline staff knowledge acquisition time from 8-12 min to 2.6 s and fault troubleshooting time from 20 min to 7 min [5]. The hazardous chemical accident root-cause prediction method RAG-Risk improves Micro-F1 and Macro-F1 by 30.16% and 48.81%, respectively, on 716 real cases [4].

© 2026 The Author(s). Authors retain copyright in their articles.

4.3 Public Services and Scientific Research

The field of public services emphasizes ease of use and inclusiveness. Xiamen University's "Law Map Assistant" builds a zero-code RAG system based on the Coze platform, integrates library laws and regulations, cases, and FAQs, and realizes second-level legal consultation responses without a GPU server, with teacher and student satisfaction reaching 90% [21]. The university feasibility study report consultation system adopts a chapter-based generation mode and combines a high-quality case weighting strategy, improving report preparation efficiency by 50% [52]. The AI4S scientific writing service framework adopts a dual-knowledge-base collaborative strategy. In the exploration stage, it calls the abstract database for rapid recall, and in the writing stage, it calls the full-text database for refined retrieval, effectively alleviating the problem of long-text loss [19]. The grass-roots grid question-answering assistant realizes the integration of "caliber query + element extraction + structured reporting" through localized policy, process, and work-order knowledge bases, reducing the reporting burden of grid workers [30]. The military equipment field constructs a multi-level knowledge graph containing 4,284 entity nodes and 4,040 relationship edges [48]. The agricultural field focuses on localized adaptation. The soil acidification diagnosis system in Chengmai, Hainan, integrates a knowledge base of soil science literature, local standards, meteorological data, and crop information, and its accuracy is 28% higher than that of a single large model [18]. Taking higher vocational Python and data structure courses as examples, a learning system integrating knowledge graphs and RAG realizes a closed loop of knowledge-point learning, intelligent question answering, and exercise practice. Knowledge retrieval accuracy reaches 94.5%, and question-answering accuracy reaches 93.2% [26]. The course question-answering system uses knowledge graph enhancement to control the hallucination rate below 5.2% and achieve answer correctness of 95.1% [25]. The aquatic nutrition recommendation system uses graphs to enforce prohibited paths, and the medical compliance rate of recommendation plans reaches 92% [37]. The travel route schematic generation framework uses a spatial semantic mapping module to automatically transform text itineraries into visual route maps and realizes personalized recommendations based on more than 70,000 scenic-spot data records and more than 1,900 question-answer pairs [49].

4.4 Security and Emergency Response: Misinformation Detection and Urban Emergency Intelligence

Highly adversarial scenarios require RAG to have robustness and interpretability. In the field of misinformation detection, RAG technology is introduced to conduct fact verification through external evidence, and multi-agent workflows are combined to simulate collaborative checking. In fake-news recognition tasks, accuracy reaches 89%, 27% higher than pure LLMs [41, 42]. The smart city emergency intelligence agent integrates vector retrieval and graph retrieval. In the case of the magnitude 6.8 earthquake in Shigatse, Tibet, information completeness improves by 35% and the hallucination rate decreases by 42% compared with a single retrieval path [53]. The community security risk disposal system HG-RAG constructs a three-layer community summary structure and extracts structured context through hierarchical retrieval. In five types of scenarios, such as theft and objects thrown from height, the generation quality score is 6.3% higher than that of traditional RAG [43].

5. Evaluation Indicator System of RAG

The evaluation of RAG systems needs to cover three dimensions: retrieval, generation, and engineering, forming a multi-indicator collaborative evaluation system. Core indicators on the retrieval side include context precision, context recall, and terminology recognition rate. Context precision measures the signal-to-noise ratio of retrieval results. In a policy consultation system, this indicator needs to be at least 0.85 to ensure that noise does not interfere with generation [15]. Context recall measures whether all relevant information required to answer a question is retrieved. It requires manually annotated knowledge fragments corresponding to the true answer for calculation. In an urban rail corpus governance system, this indicator reaches 0.92, ensuring that key information is not omitted [36]. Terminology recognition rate is a domain-specific indicator that measures the system's ability to identify professional terms. In a ship RAG system, this indicator reaches 93.7% [33]. Core indicators on the generation side include faithfulness, factual accuracy, semantic similarity, and ROUGE/BLEU indicators. Faithfulness refers to the proportion of statements in the generated answer that can be inferred from the retrieval context and reflects hallucination suppression capability. The trusted smart patent data system reaches 0.94 on this indicator through dual-path verification. Factual accuracy evaluates the

© 2026 The Author(s). Authors retain copyright in their articles.

consistency between answers and facts through manual or automatic assessment. In the lung cancer question-answering system, RAG increases this indicator for the 7B model from 32.5% to 60.0% [45]. Semantic similarity measures the semantic match between the generated answer and the true answer. In the South China Sea news question-answering benchmark, RAG increases this indicator from 0.812 to 0.865 [54]. ROUGE/BLEU, as traditional natural language processing (NLP) indicators, are used to evaluate the overlap between generated text and reference text. Core engineering indicators include response latency, provenance coverage, and robustness. Response latency refers to the time from user query to answer return. It needs to be no more than 2 seconds in policy consultation systems and no more than 1 second in edge devices such as detection instruments [15, 46]. Provenance coverage refers to the proportion of answers with valid provenance markers. The university library system reaches 100%, significantly improving user trust [22]. Robustness reflects the stability of system performance when retrieval quality fluctuates and query expressions change. DAR-RAG has an accuracy fluctuation of no more than 5% under low-quality retrieval scenarios [20].

6. Limitations and Future Trends of RAG

6.1 Existing Limitations

RAG technology currently faces five major bottlenecks. First, chunk granularity depends on manual tuning and lacks adaptive algorithms. If chunks are too large, semantics are easily mixed; if they are too small, context is easily damaged. Second, the construction cost of knowledge graphs is high, complex relation extraction still requires manual verification, and million-level entity graphs require hundreds of person-days. Third, retrieval and generation modules are optimized independently and lack end-to-end joint training, resulting in low utilization of retrieval results. Fourth, multimodal alignment technology is not mature, and complex modalities such as CAD drawings and logging curves lack unified embedding standards. Fifth, security threats continue to evolve, and new risks such as knowledge-base poisoning, post-retrieval tampering, and membership inference attacks urgently require improved protection mechanisms.

6.2 Future Development Trends

RAG will evolve in six directions. First, Agentic RAG will give autonomous planning capability to multi-agent architectures, dynamically select retrieval paths, and iteratively optimize strategies. Second, dynamic adaptive RAG will dynamically optimize retrieval parameters based on reinforcement learning and combine generation feedback to realize closed-loop optimization. Third, dual-database collaboration and trusted data frameworks will promote dual-engine modes such as “vector database + graph database” and construct bidirectional enhancement frameworks of T2S and S2T. Fourth, neural-symbolic fusion will deepen integration with graph neural networks (GNNs) and symbolic rules to improve complex reasoning capability. Fifth, edge-cloud collaborative deployment will deploy lightweight models at the edge and offload complex tasks to the cloud, realizing real-time knowledge synchronization. Sixth, vertical-domain evaluation benchmarks will be improved by constructing more domain-specific evaluation sets and promoting technical standardization.

7. Conclusion

RAG technology has evolved from the primary form of “vector database + prompt” into an engineering system of “multimodal parsing - hybrid retrieval - dual-engine reasoning - controlled generation”, becoming a key infrastructure for trustworthy AI implementation in vertical domains. This review shows that the core value of RAG does not lie in “enhancing model capability”, but in “activating domain knowledge”. Through external knowledge bases, dispersed documents, experience, and data are transformed into searchable, reasonable, and traceable “living knowledge”, effectively alleviating the knowledge hallucination, timeliness, and privacy problems of LLMs. In the future, with the development of Agentic RAG, neural-symbolic fusion, edge-cloud collaboration, and other technologies, RAG will further upgrade from an “auxiliary tool” to a “cognitive foundation” and play a core role in key fields such as complex equipment development, energy and power, and medical health. At the same time, basic issues such as data security, evaluation standards, and ethical norms need continuous attention to promote RAG technology toward greater trustworthiness, efficiency, and inclusiveness.

© 2026 The Author(s). Authors retain copyright in their articles.

References

[1] Wang, H., Xiao, X., & Huang, Z. (2026). Research on the consistency governance mechanism for enterprise system interpretation based on LLM and RAG. Communications Enterprise Management, (5), 66-72.

[2] Ren, H. Y., Sun, Z., Liao, S., Mao, D. M., Zhang, L., & Zhang, S. W. (2026). A review of the application of large language models in network threat analysis. Journal of Computer Applications, 1–26. https://doi.org/10.11772/j.issn.1001-9081.2026010138.

[3] Gu, X., Qin, S., Shu, Y., Ma, B., Liu, F., & Liu, M. A lightweight cybersecurity vulnerability risk awareness method based on RAG. Computer Science, 53(S1), 863-872.

[4] Zhang, W., Wang, Z., Li, J., He, Y., Xu, J., & Liu, Y. (2026). RAG-Risk: A large-model hazardous chemical accident root-cause prediction method based on RAG. Journal of Systems Science and Mathematical Sciences, 1-21. https://link.cnki.net/urlid/11.2019.O1.20260526.1024.014

[5] Luo, J. (2026). Design and engineering application of a railway four-electricity knowledge-enhanced retrieval system based on the RAG architecture. Modern Information Technology, 10(11), 73-80. https://doi.org/10.19850/j.cnki.2096-4706.2026.11.014

[6] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, 9459-9474.

[7] Dai, R. (2026). Research method for an MBSE data foundation based on RAG-large language models. Modern Navigation, 17(3), 191-196.

[8] Wang, H., Hou, Q., Fang, Z., Jiang, H., & Shen, W. (2026). Construction and application of an intelligent knowledge system for offshore oilfield logging based on improved Retrieval-Augmented Generation (RAG). Mud Logging Engineering, 37(2), 1-8.

[9] Xue, Z., Zhang, Y., Zheng, Z., Li, C., & Jiang, H. (2026). Intelligent power retrieval engine based on knowledge graph path reasoning and large models. Power System Technology, 50(8), 3263-3280. https://doi.org/10.13335/j.1000-3673.pst.2025.1205

[10] Yao, M. (2026). Progress of RAG and its application in monitoring systems. China Computer & Communication, 38(14), 133-135.

[11] Guo, L., Chen, J., & Li, J. (2026). Research on a hybrid general document parsing method for RAG. Software Guide, 1-7. https://link.cnki.net/urlid/42.1671.TP.20260721.1641.020

[12] Zhang, Y. (2026). Research on an intelligent recipe recommendation system based on Retrieval-Augmented Generation. China Computer & Communication, 38(16), 86-88.

[13] Feng, Y., Xue, S., Wang, S., Shu, H., Zhang, W., & Li, Q. (2026). Intelligent chunking of design documents based on layout parsing and LLM. Electric Power Survey & Design, (7), 84-88. https://doi.org/10.13500/j.dlkcsj.issn1671-9913.2026.07.015

[14] Xue, H., Zhang, Z., Liang, W., & Wang, J. (2026). Research on an LLM semantic communication knowledge synchronization method based on RAG retrieval enhancement. China-Arab States Science and Technology Forum (Chinese and English), (7), 81-86.

[15] Chi, Z., Pan, X., Huang, X., Zhou, X., & Huang, X. (2026). Optimization of a RAG system in policy consultation based on hybrid vector and tag retrieval. Industrial Control Computer, 39(7), 96-97, 100.

[16] Zhang, Q., Qu, P., & Zuo, L. (2026). Intelligent question-answering system based on large language models and improved RAG. Computer Systems & Applications, 1-10. https://doi.org/10.15888/j.cnki.csa.010297

[17] Xie, B., He, N., Zhang, R., & Yang, Y. (2026). RAG-SymbolicDet: Visual event symbolic reasoning with multi-strategy Retrieval-Augmented Generation. Computer Engineering and Applications, 1–12. https://link.cnki.net/urlid/11.2127.tp.20260727.1343.004

© 2026 The Author(s). Authors retain copyright in their articles.

[18] Zhang, X., Chen, S., Xie, F., Hu, Y., Yang, H., Wang, L., & Xie, Q. (2026). A soil acidification obstacle diagnosis system based on Retrieval-Augmented Generation (RAG) coordinated with large models: A case study of Chengmai, Hainan. Journal of Agro-Environment Science, 1–25.

[19] Wang, Y., Sun, L., & Fu, Z. (2026). Research on scientific writing knowledge services integrating dual-database collaborative RAG under the AI4S paradigm: A case of AI-domain scientific writing. Journal of Modern Information, 46(8), 3-16.

[20] He, Q., Li, Q., Guo, H., & Chen, Q. (2026). Research on a retrieval-augmented generation method based on dynamic adaptive reinforcement learning. Journal of Frontiers of Computer Science and Technology, 1-15. https://link.cnki.net/urlid/11.5602.tp.20260727.1223.006

[21] Xu, G. (2026). Practice and enlightenment of intelligent question answering with large language models driven by the Coze + RAG architecture in libraries: A case study of Xiamen University's "Law Map Assistant". Value Engineering, 45(20), 165-168.

[22] Pan, X. (2026). Construction and application of an intelligent reference consultation system for university libraries based on the RAG architecture. Journal of Academic Library and Information Science, 44(4), 15-21.

[23] Huang, Z., & Qiao, X. (2026). Construction and application of a classroom intelligent assistant based on AIGC and RAG technology. Electronic Component and Information Technology, 10(5), 216-220. https://doi.org/10.19772/j.cnki.2096-4455.2026.05.061

[24] Wu, X., & Pei, C. (2026). Research on an intelligent tutoring system based on large-model RAG technology. China-Arab States Science and Technology Forum (Chinese and English), (6), 92-96.

[25] Lin, S. (2026). Research on a course intelligent question-answering system integrating knowledge graphs and RAG. Fujian Computer, 42(5), 68-73. https://doi.org/10.16707/j.cnki.fjpc.2026.05.012

[26] Liu, Y., Dong, S., Zhang, S., Cheng, Y., & Zhang, S. (2026). Design and implementation of an enhanced large language model learning system coordinating RAG and knowledge graphs: A case study of higher vocational Python and data structure courses. Computer Knowledge and Technology, 22(15), 137-140. https://doi.org/10.14004/j.cnki.ckt.2026.0697

[27] Wang, J., & Ma, B. (2026). Application of Graph-RAG technology based on structural semantic collaborative optimization in intelligent inquiry systems for equipment operation and maintenance. Journal of Beijing University of Chemical Technology (Natural Science Edition), 1-10. https://link.cnki.net/urlid/11.4755.TQ.20260602.1102.002

[28] Liu, X., Xu, L., Ma, H., & Liu, S. (2026). Research on RAG question-answering optimization for fault diagnosis of coal mine electromechanical equipment. Coal Technology, 45(6), 139-143. https://doi.org/10.13301/j.cnki.ct.2026.06.026

[29] Xiong, L., Du, S., Shi, H., Hu, J., Yang, Y., & Li, T. (2026). Government hotline question-answering system based on a knowledge-enhanced large language model architecture. Journal of Computer Applications, 46(6), 1721-1727.

[30] Wang, Y., Liu, D., Che, S., & Wei, Y. (2026). Technical research on a grass-roots grid question-answering assistant based on large models and RAG. Jiangsu Communications, 42(3), 19-24, 50.

[31] Zhao, J. (2026). Research on a BIM knowledge graph intelligent retrieval system based on Graph-RAG. Modern Information Technology, 10(11), 62-67. https://doi.org/10.19850/j.cnki.2096-4706.2026.11.012

[32] Yang, Y., Yang, C., Wu, H., Zhu, G., Liu, M., & Ren, Z. (2026). Hierarchical criterion-driven profile construction and graph retrieval-augmented bidding screening method. Journal of Frontiers of Computer Science and Technology, 1–17.

[33] Zhang, Y., Liu, Y., Li, L., Qi, D., Xu, D., Li, Y., & Wei, H. (2026). A ship RAG system enhanced by ship professional vocabulary. Journal of Shanghai Ship and Shipping Research Institute, 49(3), 33-39, 62.

© 2026 The Author(s). Authors retain copyright in their articles.

[34] Sun, Y., Wang, Z., Liu, H., Wang, W., & Jiang, M. (2026). Construction of an intelligent question-answering system for Python programming based on LLM and Advanced RAG. Research and Exploration in Laboratory, 45(8), 104-110, 116. https://doi.org/10.19927/j.cnki.syyt.2026.08.015

[35] Li, Z., Jiang, D., & Wang, H. (2026). Exploration and design of a university knowledge question-answering system based on large models and RAG technology. China Computer & Communication, 38(12), 41-43.

[36] Wang, C., Wang, H., Zhang, B., Zhou, M., & Xu, Y. (2026). Research and practice of corpus governance for an urban rail quality and safety large model driven by RAG. Tunnel Construction, 1-11. https://link.cnki.net/urlid/44.1745.U.20260727.1033.002

[37] Gao, J., Ai, X., & Zhang, S. (2026). LLM aquatic nutrition recommendation architecture based on Hybrid RAG. Journal of Shanghai Ocean University, 35(4), 855-866.

[38] Dou, J., Wang, X., Qiao, J., Jin, M., Li, J., Li, Y., & Han, Y. (2026). Distribution network maintenance intelligent auxiliary decision-making framework based on knowledge graph Retrieval-Augmented Generation and large language models. Advanced Engineering Sciences, 58(4), 67-79.

[39] Shen, J., Chen, H., & Chen, X. (2026). Multi-RAG: A distributed Retrieval-Augmented Generation framework for cross-domain data. Computer Science, 53(S1), 472-478.

[40] Xu, Y., Ni, H., Jiang, M., & Mei, Z. (2026). Research on an intelligent auxiliary design method for ship pipeline systems based on RAG and knowledge graphs. Ship Materials & Market, 34(6), 16-19. https://doi.org/10.19727/j.cnki.cbwzysc.2026.06.006

[41] Zhang, L., Huang, C., Ding, W., Pan, Y., Zhu, Z., & Yang, H. (2026). Intelligent transformation of misinformation detection: Technical system, dilemmas, and prospects based on large language models. Information Security and Communications Privacy, (7), 9-25.

[42] Wang, Y., Dai, S., Li, J., Wang, G., & Sun, J. (2026). A review of the application of Retrieval-Augmented Generation in misinformation detection. Computer Engineering and Applications, 1-33. https://link.cnki.net/urlid/11.2127.TP.20260804.1915.006

[43] Jing, Z., Hu, X., & Hu, J. (2026). Generation of community security risk disposal recommendations based on hierarchical graph retrieval augmentation. Journal of Computer Applications, 1-13. https://link.cnki.net/urlid/51.1307.TP.20260723.1744.004

[44] Feng, Z., Zhang, Y., Yang, X., & Chen, T. (2026). Theoretical construction and framework implementation of trusted smart data for patent literature. Library Tribune, 1-16. https://link.cnki.net/urlid/44.1306.G2.20260720.1026.002

[45] Ren, Z., Li, Y., Ma, Y., Chen, Q., Li, J., Yang, S., & Liu, L. (2026). Efficacy evaluation of domestic lightweight large language models integrated with Retrieval-Augmented Generation for lung cancer specialty question answering. Chinese Journal of Clinical Thoracic and Cardiovascular Surgery, 33(9), 1403-1411.

[46] Li, Y., Li, D., Meng, X., Liu, X., & Cui, H. (2026). Research on localized detection instruments based on AI agents. Automation & Instrumentation, 41(7), 106-109, 114. https://doi.org/10.19557/j.cnki.1001-9944.2026.07.023

[47] Tang, L., Li, W., Huang, P., Huang, X., Wang, Q., & Liu, J. (2026). Intelligent recommendation system for Chinese patent medicine based on cloud services and RAG technology. Computer Science, 53(S1), 213-219.

[48] Yue, C. (2026). Research on a military equipment domain knowledge question-answering system integrating knowledge graphs and large language models [Doctoral dissertation, Yantai University]. https://doi.org/10.27437/d.cnki.gytdu.2026.000149

[49] Sun, C., Lan, T., Wu, Z., Cheng, D., Shi, X., & Jiang, S. (2026). Travel route schematic generation framework based on LLM and RAG. Surveying and Mapping, 49(3), 169-174.

© 2026 The Author(s). Authors retain copyright in their articles.

[50] Li, M., Zhang, C., Hu, H., Huang, L., & Zhang, Z. (2026). Auxiliary application of large language models in finite element modeling of ship structures. Journal of Shanghai Ship and Shipping Research Institute, 49(3), 27-32.

[51] Li, W., Zhang, Z., Tao, W., Liu, C., Wang, L., & Li, R. (2026). Intelligent frequency regulation of power systems assisted by large language models. High Voltage Engineering, 52(7), 3010-3021. https://doi.org/10.13336/j.1003-6520.hve.20250201

[52] Liu, J., Zhou, Q., Guo, R., & Feng, J. (2026). Applied research based on large language models and Retrieval-Augmented Generation: A case study of a feasibility study report consultation system. China Engineering Consulting, (7), 97-105.

[53] Zhao, Y., Huang, W., & Liu, Y. (2026). Construction of a smart city emergency intelligence service agent integrating large models and Hybrid-RAG. Journal of Modern Information, 1-14. https://link.cnki.net/urlid/22.1182.G3.20260706.1521.004

[54] Feng, S., & Yang, H. (2026). Construction and empirical study of a Retrieval-Augmented Generation evaluation benchmark based on South China Sea news dynamics. Journal of Intelligence, 1-8. https://link.cnki.net/urlid/61.1167.g3.20260728.1418.010

Funding

The authors gratefully acknowledge the financial support provided by the 2026 Science and Technology Innovation Think Tank Project of the Liaoning Provincial Association for Science and Technology (No. LNKX2026QN62).

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Copyrights

Copyright for this article is retained by the author (s), with first publication rights granted to the journal. This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).

© 2026 The Author(s). Authors retain copyright in their articles.