

Retrieval-Augmented Generation Applications and Development in the Field of Electronic Test Instruments

Jie Hu^{1, 2, *}, Yuanjun Li^{1, 2}, Chunhui Li^{1, 2}, Lingzi Chen^{1, 2}, Fei Li^{1, 2}

¹ National Key Laboratory of Science and Technology on Electronic Test and Measurement, Qingdao, China

² The 41st Institute of China Electronic Technology Group Corporation, Qingdao, China

* Corresponding author: Jie Hu

Abstract: The rapid development of artificial intelligence (AI), especially the fast iteration and widespread deployment of large language models (LLMs), has ushered in a new paradigm for the intelligent augmentation of electronic test instruments and automatic test systems (ATS). By combining the powerful reasoning and generative capabilities of LLMs with the timeliness and accuracy of external knowledge bases, retrieval-augmented generation (RAG) has emerged as a highly promising approach to addressing knowledge-intensive challenges in the operation and maintenance of electronic test instruments. This paper presents a comprehensive review of RAG and its applications in the field of electronic test instruments. First, the basic architecture of RAG is reviewed, including retrieval mechanisms, knowledge-base construction, and generative models. Then, considering the operational characteristics of electronic test instruments and automatic test systems, key knowledge-intensive tasks are analyzed, such as fault diagnosis, test configuration, and process planning, followed by a discussion of representative RAG-based applications, including intelligent fault diagnosis systems, adaptive test-parameter configuration, and interactive question-answering assistants. In addition, practical considerations for engineering implementation are discussed with respect to knowledge-base construction and effectiveness evaluation. Finally, open issues and future research directions are outlined, including multimodal RAG, federated learning for distributed knowledge bases, and edge deployment. This review aims to provide researchers and practitioners in relevant industries with a comprehensive understanding of the current status and development prospects of RAG technologies in electronic test instruments.

Keywords: Retrieval-augmented generation, Electronic test instruments, Knowledge base, Intelligent fault diagnosis, Large language models, Test automation.

1. Introduction

Electronic test instruments, such as signal generators, oscilloscopes, spectrum analyzers, and vector network analyzers, form the cornerstone of modern electronic engineering. They are indispensable to research, development, production, and maintenance in a wide range of industries, including communications, radar, aerospace, automotive, and semiconductor testing. With the emergence of high-frequency, high-bandwidth, and multi-standard technologies, the complexity of these instruments has increased exponentially, giving rise to ever more sophisticated operating procedures and increasingly extensive technical documentation.

Traditional methods for operating test instruments rely primarily on manually consulting lengthy technical manuals, drawing upon expert experience, and adjusting parameters through repeated trial and error. These approaches suffer from several limitations [1, 2]: (1) they impose high requirements on operators' professional competence and demand substantial domain-specific expertise; (2) fault diagnosis and resolution are often time-consuming; (3) knowledge is difficult to reuse across different types of instruments and task scenarios; (4) scalability is limited when dealing with complex test systems integrated from multiple types of instruments.

Meanwhile, with the continuous development of LLMs [3], such as Qwen, DeepSeek, GPT, and Claude, as well as their open-source counterparts, natural language processing has advanced significantly. Although LLMs demonstrate remarkable capabilities in text understanding and content generation, they exhibit inherent limitations when applied to

domain-specific tasks involving specialized business logic: (1) the knowledge cutoff problem, LLMs cannot access information beyond their training data; (2) hallucinations, LLMs may generate technically plausible but incorrect information; and (3) insufficient specialization, general LLMs often lack in-depth domain knowledge related to electronic test instruments.

RAG is a technical paradigm that combines pretrained parametric language models with non-parametric retrieval systems. Its primary objective is to enhance and guide the text generation process of LLMs by precisely retrieving relevant information fragments from large-scale corpora, making it particularly suitable for knowledge-intensive tasks with stringent requirements on factual accuracy and timeliness. By introducing a retrieval mechanism, RAG provides the generative model with fact-based anchors, thereby significantly mitigating the risk of hallucinations while maintaining the fluency of generated responses.

In the field of electronic test instruments, RAG holds transformative application potential. By constructing a comprehensive knowledge base that incorporates instrument specifications, fault case repositories, test procedures, and expert knowledge, RAG can provide context-aware, real-time guidance for diverse task scenarios and deliver intelligent decision support for operators. This paper aims to present an overview of RAG technology and its applications in electronic test instruments, covering its architectural foundations, application paradigms, practical considerations, and future research directions.

The remainder of this paper is organized as follows. Section 2 reviews the fundamental concepts and architecture

of RAG. Section 3 examines the domain characteristics of electronic test instruments and identifies key knowledge-intensive tasks. Section 4 elaborates on RAG-based application paradigms in detail. Section 5 discusses relevant considerations for implementation. Section 6 analyzes the challenges and future development directions.

2. Fundamentals of Retrieval-Augmented Generation

2.1. RAG Architecture

RAG can be decomposed into the coordinated operation of three key components: a knowledge-base indexer, which processes and stores documents in a retrievable format, a retriever, which accesses the “non-parametric knowledge” contained in the knowledge base, and a generator, which reasons over the user’s original question together with the external knowledge fragments provided by the retriever and produces a response.

The knowledge-base indexer typically employs a document preprocessing pipeline that includes text extraction, chunking, embedding generation, and vector database indexing. Modern RAG systems use dense retrieval methods based on neural embedding models, such as dense passage retrieval [4] or approximate nearest neighbor negative contrastive learning [5] to encode documents into dense vector representations.

Upon receiving a user input, the retriever is responsible for rapidly locating and extracting the most relevant information fragments from a large-scale knowledge base. It commonly adopts a two-stage retrieval process: coarse-grained candidate retrieval and fine-grained reranking. The former is performed using approximate nearest neighbor (ANN) search algorithms, such as HNSW (Hierarchical Navigable Small World) or IVF (Inverted File Index), and the latter is conducted by computing attention-based relevance scores between each candidate document and the query using a cross-encoder model.

The generator is usually instantiated as a decoder-only transformer model. It receives an augmented input composed of the original query and the retrieved passages, and is required to use the retrieved external knowledge as the primary factual basis while combining it with its own parametric knowledge to reason, integrate, and refine the final output, thereby generating a coherent response that is more accurate, information-rich, and faithful to external facts. Recent advances in instruction tuning and chain-of-thought prompting have significantly improved the reasoning capability of generator in RAG systems [6].

2.2. Retrieval Mechanisms

The retriever can adapt its retrieval strategy according to different retrieval granularities and the application scenarios. Keyword-based retrieval [7] adopts traditional BM25 or TF-IDF scoring methods and is computationally efficient, however, it inherently suffers from the vocabulary mismatch problem and struggles to capture semantic relationships. Semantic retrieval uses embedding models to map text into a continuous vector space, enabling the capture of semantic similarity beyond lexical overlap, it is therefore particularly suitable for retrieving technical documentation where terminological variants are common [5, 8]. Hybrid retrieval integrates keyword-based retrieval results with semantic retrieval scores through methods such as reciprocal rank

fusion (RRF) or learned rank aggregation, hybrid methods have consistently outperformed retrieval systems based on a single strategy in practical evaluations [9]. Graph-based retrieval leverages knowledge graphs to perform multi-hop reasoning over structured knowledge, such methods excel at multi-hop retrieval and deep reasoning, and are therefore widely applied in complex fault diagnosis scenarios [10].

2.3. Generation Models

The generator in RAG architecture has evolved from simple sequence-to-sequence models to large-scale pretrained language models. Recent advances also include instruction-tuned models fine-tuned on instruction-following datasets to better satisfy user requirements [11], chain-of-thought prompting [6], which guides models to generate intermediate reasoning steps before producing the final answer, and self-correction mechanisms [12], which enable models to verify and refine their outputs by retrieving additional information or performing consistency checks.

3. Electronic Test Instruments: Domain Characteristics

3.1. Classification and Applications

Electronic test instruments serve as the “eyes” and “yardstick” of the electronic information industry, fulfilling the core mission of testing, analyzing, and verifying signals, electronic components, and system performance. These instruments can be divided by test function into five major categories. The first category comprises basic electrical-parameter measurement instruments, such as digital multimeters and LCR testers, which are used to measure fundamental electrical quantities including voltage, current, resistance, capacitance, and inductance. These are the standard tools for engineers’ daily debugging work. The second category includes signal analysis and waveform observation instruments, such as oscilloscopes, spectrum analyzers, vector network analyzers, and noise figure analyzers, which are used to observe time-domain waveforms, spectral characteristics, network parameters, and digital timing, and constitute the core means for circuit design and fault troubleshooting. The third category consists of power supply and load test equipment, such as programmable DC power supplies, electronic loads, and microwave power meters, which are used to provide stable voltage/current sources, simulate load conditions, and measure power quality. The fourth category covers safety and environmental reliability test equipment, including withstand voltage testers, insulation resistance testers, and temperature-humidity test chambers, which are used to verify product safety and durability. The fifth category refers to test system with ever-increasing levels of automation, which is commonly used in large-volume automated inspection of semiconductor chips, printed circuit board assemblies, and new energy batteries to perform functional verification and performance screening.

Electronic test instruments typically span the entire lifecycle of electronic products, including design, manufacturing, and after-sales maintenance. In the design stage, instruments such as oscilloscopes, spectrum analyzers, and logic analyzers are used for prototype circuit debugging, signal integrity analysis, and communication protocol verification, helping engineers rapidly locate design defects. On the production line, ATS, LCR testers, and safety testers form the core capability, ensuring consistency and reliability

of outgoing product quality. In particular, in fields such as switching power supplies, battery management systems for new energy vehicles, and consumer electronics, where automated testing is critical to throughput. Various specialized test instruments are widely used in satellite, communications, navigation, radar, scientific research, and education, where they play an essential role in providing test and measurement support.

3.2. Knowledge-Intensive Tasks

The operation and maintenance of electronic test instruments encompass a variety of knowledge-intensive tasks, including fault diagnosis, test configuration, and test process planning. These tasks are, in essence, highly constrained, knowledge-heterogeneous, and strongly procedural.

Fault diagnosis requires associating fault symptoms with potential causes, a process that often depends on large fault-case databases, maintenance manuals, and expert experience. The diagnostic workflow usually involves multi-step reasoning and cross-document analysis of complex business logic and technical details of equipment operation and maintenance support.

Test configuration requires selecting and setting appropriate instrument parameters, such as frequency range, resolution bandwidth, filter type, and triggering, for specific test scenarios while taking timing relationships into account.

This requires a deep understanding of both the capabilities of the device under test and those of the instrument, unreasonable configuration may lead to measurement errors or missed defects [2].

The purpose of test process planning is to transform abstract test requirements into executable, measurable, and controllable systematic action plans. The planning process incorporates instrument manuals, specification standards, and historical test data, and covers all necessary measurements, verifications, and compliance checks. Retrieval of technical documentation requires operators to locate specific information from lengthy instrument manuals, application notes, and specification standards.

In addressing the above tasks, relying solely on on-site test personnel is complex, time-consuming, and costly in trial and error, while traditional large models are prone to outdated knowledge, device blind spots, and unreliable workflows. RAG alleviates these problems through a “retrieve-then-generate” paradigm by unifying instrument manuals, calibration specifications, historical fault tickets, SCPI command sets, test configuration templates, automation test scripts, and real-time logs into a knowledge service that is “retrievable, traceable, and executable,” making it well suited to such knowledge-intensive tasks. The mapping between RAG capabilities and test instrument tasks is shown in Table 1.

Table 1. Mapping between RAG capabilities and test instrument tasks

RAG capability	Test instrument task	Knowledge source	Effect
semantic retrieval	fault symptom matching	fault-case database	improved diagnostic speed
context-aware generation	test configuration recommendation	instrument manual	reduced parameter configuration time
multi-hop reasoning	root cause analysis	maintenance manual, logs	reduced repeat repairs
instruction following	test process planning	standards, specifications	improved test efficiency
conversational interface	operator assistance	knowledge base	improved retrieval efficiency

4. Typical Applications Based on Retrieval-Augmented Generation

4.1. Intelligent Fault Diagnosis

Intelligent fault diagnosis is one of the most influential representative applications of RAG in the field of electronic test instruments. The diagnostic workflow typically follows a structured sequence of operations: symptom collection, fault isolation, root-cause localization, and maintenance recommendation [13].

The workflow of a RAG-based fault diagnosis system is as follows. First, the operator describes the fault phenomenon of the instrument in natural language—for example, “when measuring low-level signals, the spectrum analyzer exhibits an elevated noise floor above 5 GHz.” The retrieval module then searches the fault-diagnosis knowledge base, which focuses on fault-case repositories and maintenance manuals, and returns the top-k historical cases most similar to the current fault together with their corresponding troubleshooting steps. Subsequently, the generation module synthesizes this information and produces a diagnostic report containing prioritized fault hypotheses, recommended diagnostic procedures, and maintenance instructions.

The fault-diagnosis knowledge base usually includes: (1) structured fault-code databases maintained by instrument manufacturers, (2) instrument user manuals, maintenance

manuals, and fault cases, (3) unstructured service logs, and (4) community-contributed troubleshooting guides.

4.2. Adaptive Test Parameter Configuration

Appropriately matched test parameter configuration on the instrument panel is crucial for obtaining accurate measurement results in a timely manner. Improper configuration may introduce measurement errors, such as those caused by improper input range selection, insufficient number of averages, erroneous analysis-bandwidth or trigger settings, leading to invalid measurements, higher trial-and-error costs, and degraded test throughput.

A RAG-based adaptive test parameter configuration assistant operates as follows: (1) it understands the user’s test objective through natural-language description and confirms the test task, for example, “I need to measure the harmonic distortion of a 2.4 GHz Wi-Fi transmitter, and its spurious level should be below -40dBc”, (2) it retrieves relevant configuration guidelines from instrument technical documentation and a rule knowledge base, and (3) it generates a step-by-step configuration procedure and parameter key-value pairs tailored to the specific instrument model and test scenario.

The rule knowledge base is a multidimensional parameter space in which each dimension corresponds to a programmable instrument parameter, such as frequency range, reference level, sampling rate, resolution bandwidth, video

bandwidth, detector type, and sweep time. The adaptive configuration process generates a parameter configuration scheme suited to the test task through test-task parsing, case-base retrieval, and rule knowledge-base matching.

4.3. Interactive Question-Answering Assistant

A more general and widely adopted application of RAG is intelligent question answering over domain knowledge. In practice, this usually takes the form of customized “interactive question-answering assistants” for various vertical industries, which help users resolve knowledge-service problems such as real-time feedback, information retrieval, and pre-sales/post-sales service through client-side dialog interfaces, thereby significantly reducing communication costs and improving service efficiency and quality.

In interactive question answering, RAG has evolved from the initial stage of “stuffing documents into a vector database and letting the model answer” into a composite capability encompassing hybrid retrieval, reranking, multi-turn clarification, traceable generation, and agent-based tool invocation. It is no longer limited to single-turn question answering, instead, it can select retrieval strategies according to question complexity, ask the user clarifying questions when necessary, decompose problems into sub-questions, invoke tools, and annotate the sources of answers. Industrial deployment is already relatively mature, with scenarios in finance, healthcare, law, manufacturing, customer service, and operations and maintenance all adopting RAG for intelligent question answering and agent assistance. For example, customer-service scenarios can implement multi-turn question answering based on product manuals, historical tickets, and policy documents, healthcare and legal scenarios emphasize traceability and source attribution, and manufacturing and operations-and-maintenance scenarios are beginning to connect equipment manuals, fault cases, and real-time status to RAG, enabling frontline operators to troubleshoot problems using natural language.

Conversational question answering ensures the coherence of multi-turn interaction through dialogue state tracking and context management. Retrieval is performed in every turn and combined with dialogue history to resolve coreference and maintain context. This approach significantly improves operators’ productivity and satisfaction in field-operation scenarios [14, 15].

5. Practical Considerations

5.1. Domain Knowledge-Base Construction

The effectiveness of RAG applications in vertical domains depends largely on high-quality domain knowledge bases, and the construction of knowledge bases for electronic test instruments is no exception. This process requires careful consideration of multiple factors. Data sources should integrate manufacturer documentation (user manuals, maintenance manuals, and application notes), test standards (IEEE, IEC, and MIL-STD), historical test data, fault-case records, and experiential knowledge accumulated by experts [12, 16, 17].

The chunking strategy is of critical importance, document chunking must strike a balance between retrieval granularity and context preservation. For technical manuals, chunking by logical sections, such as by chapter or functional module, often outperforms fixed-length text chunking [18]. By

attaching structured metadata to document chunks (e.g., instrument model, measurement type, frequency range, and error code), hybrid retrieval combining semantic search with metadata filtering can be achieved.

Semantic vector libraries constructed through word embedding models have inherent limitations. Traditional RAG is not well suited to pure logical reasoning, strong real-time control, highly deterministic rule-based judgment, or complex statistical computation. A domain ontology can be used to construct an entity–relation network, and a knowledge graph for electronic test instruments can be built in a business-oriented manner to better support multi-hop reasoning, rule validation, high-certainty inference, and hybrid retrieval requirements [15, 16, 19].

5.2. Effectiveness Evaluation

When evaluating RAG application systems in the field of electronic test instruments, domain-specific metrics are required rather than relying solely on standard information-retrieval measures. Effectiveness evaluation should shift from assessing “whether the answer looks plausible” to assessing “whether it can be executed.” Evaluation should therefore not focus only on whether the response is fluent, but also on whether it identifies the correct instrument model, options, test items, and matching test methods, whether safety constraints are omitted, whether the generated scripts or configuration schemes are executable, and whether the citations are authentic and traceable.

6. Challenges and Future Directions

Although electronic test instruments continue to advance in integration, networking, and intelligence, the deployment and application of RAG technology in electronic test instruments or related test systems still faces numerous challenges.

Knowledge timeliness remains an urgent problem to be addressed. As instrument firmware and test standards evolve rapidly, a mechanism for continuous knowledge-base updating must be established. Federated learning methods, which enable distributed knowledge contribution without centrally storing sensitive test data, are regarded as a promising research direction [20–22].

Multimodal fusion presents another challenge. In real-world instrument diagnosis, multiple information modalities must often be processed simultaneously, including textual documents, instrument status screenshots, waveform and spectrum images, and log files. Multimodal RAG systems capable of joint retrieval and reasoning over heterogeneous data types have emerged as an important research frontier [23].

Explainability and trust are paramount in safety-critical test applications. In aerospace, medical devices, and mission-critical test support, the attribution of retrieved evidence as well as the provenance of both the testing process and its results must be verifiable and traceable to authoritative sources [24, 25].

Edge deployment presents both opportunities and challenges. In field service scenarios where network connectivity is often limited or intermittent, running a RAG system on an edge device with a compressed model and an on-device knowledge base calls for breakthroughs in model compression, quantization, and retrieval efficiency under tight hardware constraints [26–28].

Acknowledgements

This research was supported by research fund No. 2024WDZC202603.

References

- [1] Xie, Y., He, K., & Castellanos-Gomez, A. (2025). Toward full autonomous laboratory instrumentation control with large language models. *Small Structures*, 6(8). <https://doi.org/10.1002/sstr.202500173>
- [2] Zhang, Y., Liu, J., Zhang, Z., Qi, L., Wang, P., Wang, X., & Zhao, W. (2026). LLM-enabled multi-agent collaborative instrument control for automated microelectronic testing. *Advanced Devices & Instrumentation*, 7, 0202. <https://spj.science.org/doi/10.34133/adi.0202>
- [3] Peykani, P., Ramezanlou, F., Tanasescu, C., & Ghanidel, S. (2025). Large language models: A structured taxonomy and review of challenges, limitations, solutions, and future directions. *Applied Sciences*, 15(14), 8103. <https://doi.org/10.3390/app15148103>
- [4] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W., Rocktaschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33. <https://doi.org/10.48550/arXiv.2005.11401>
- [5] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781. <https://doi.org/10.18653/v1/2020.emnlp-main.550>
- [6] Trivedi, H., Balasubramanian, N., Khot, T., & Sabharwal, A. (2023). Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 1, 10014–10037. <https://doi.org/10.18653/v1/2023.acl-long.557>
- [7] Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333–389. <https://doi.org/10.1561/15000000019>
- [8] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. <https://doi.org/10.18653/v1/d19-1410>
- [9] Cormack, G. V., Clarke, C. L. A., & Buettcher, S. (2009). Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 758–759. <https://doi.org/10.1145/1571941.1572114>
- [10] Liu, P., Qian, L., Zhao, X., & Tao, B. (2024). Joint knowledge graph and large language model for fault diagnosis and its application in aviation assembly. *IEEE Transactions on Industrial Informatics*, 20(6), 8160–8169. <https://doi.org/10.1109/tii.2024.3366977>
- [11] Mao, K., Liu, Z., Qian, H., Mo, F., Deng, C., & Dou, Z. (2024). RAG-Studio: Towards in-domain adaptation of retrieval augmented generation through self-alignment. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 725–735. <https://doi.org/10.18653/v1/2024.findings-emnlp.41>
- [12] Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *ICLR* 2024. <https://doi.org/10.48550/arxiv.2310.11511>
- [13] Liu, P., Qian, L., Zhao, X., & Tao, B. (2024). Joint knowledge graph and large language model for fault diagnosis and its application in aviation assembly. *IEEE Transactions on Industrial Informatics*, 20(6), 8160–8169. <https://doi.org/10.1109/tii.2024.3366977>
- [14] Mo, F., Mao, K., Zhao, Z., Qian, H., Chen, H., Cheng, Y., Li, X., Zhu, Y., Dou, Z., & Nie, J. (2025). A survey of conversational search. *ACM Transactions on Information Systems*, 43(6). <https://doi.org/10.1145/3759453>
- [15] Picariello, F., Daponte, P., Vito, L., & Tudosa, I. (2026). AI-supported remote laboratories for measurement education: A preliminary evaluation of LLM for oscilloscope-based training. *2026 IEEE International Workshop on Metrology for Industry 40 & IoT (MetroInd40 & IoT)*, 662–667. <https://doi.org/10.1109/metroind40.0iot69397.2026.11653169>
- [16] Huang, L., Cui, Q., Yu, L., & Dai, J. (2026). Dynamic knowledge graph construction from multi-source heterogeneous data for RAG-based vertical domain question answering. *2026 5th International Conference on Artificial Intelligence, Internet of Things and Cloud Computing Technology (AIoTC)*, 117–122. <https://doi.org/10.1109/aiotc71250.2026.11688541>
- [17] Siddharth, L., & Luo, J. (2024). Retrieval augmented generation using engineering design knowledge. *Knowledge-Based Systems*, 303, 112410. <https://doi.org/10.1016/j.knosys.2024.112410>
- [18] Mattos, E., Nascimento, E., Amoedo, D., Guimarães, V., Matias, P., Oliveira, E., Naja, L., Castro, M., Oliveira, P., Uchoa, M., Maia, C., Botinelly, L., Brito, R., Cardoso, B., Leite, L., Pessoa, L., Pires, A., Miranda, A., Pimentel, A., Belém, R., Fabricio, R., Carvalho, C., & Júnior, W. (2026). Chunking strategies in RAG pipelines: Evaluating the impact on LLM response quality using consumer electronics user manuals. *2026 IEEE International Conference on Consumer Electronics (ICCE)*, 1–6. <https://doi.org/10.1109/icce67443.2026.11449886>
- [19] Chen, X., Jia, S., & Xiang, Y. (2020). A review: Knowledge reasoning over knowledge graph. *Expert Systems with Applications*, 141. <https://doi.org/10.1016/j.eswa.2019.112948>
- [20] Yang, R., Wong, M., Li, H., Li, X., Zhu, W., Liao, J., Yu, K., Liew, J., Xuan, W., Chen, Y., Ke, Y., Ong, J., Teodoro, D., Hong, C., Ting, D., & Liu, N. (2026). Retrieval-augmented generation in medicine: A scoping review of technical implementations, clinical applications, and ethical considerations. *Cell Reports Medicine*. <https://doi.org/10.1016/j.xcrm.2026.102927>
- [21] Li, Z., Wang, Z., Wang, W., Hung, K., Xie, H., & Wang, F. (2025). Retrieval-augmented generation for educational application: A systematic survey. *Computers and Education: Artificial Intelligence*, 8, 100417. <https://doi.org/10.1016/j.caeai.2025.100417>
- [22] Brown, A., Roman, M., & Devereux, B. (2025). A systematic literature review of retrieval-augmented generation: Techniques, metrics, and challenges. *Big Data and Cognitive Computing*, 9(12), 320. <https://doi.org/10.3390/bdcc9120320>
- [23] Wang, R., Wang, J., Huang, Y., Guo, H., & Pang, J. (2026). A survey on medical multimodal retrieval-augmented generation (RAG). *High-Confidence Computing*. <https://doi.org/10.1016/j.hcc.2026.100415>
- [24] Nandhakumar, K. P., & Sasikala, E. (2026). Intelligent ERP AI agent with RAG and LLM integration: A systematic literature review. *2026 4th International Conference on Artificial Intelligence and Machine Learning Applications (AIMLA)*, 1–7. <https://doi.org/10.1109/AIMLA67915.2026.11522318>

- [25] Lambert, B., Forbes, F., Doyle, S., Dehaene, H., & Dojat, M. (2024). Trustworthy clinical AI solutions: A unified review of uncertainty quantification in deep learning models for medical image analysis. *Artificial Intelligence in Medicine*, 150, 102830. <https://doi.org/10.1016/j.artmed.2024.102830>
- [26] Zhao, W., Yang, X., Lyu, Z., Xu, C., & Guan, Z. (2025). Road of large language model: Source, challenge, and future perspectives. *Research*, 8, 0655. <https://doi.org/10.34133/research.0655>
- [27] Bianchini, F., Bianchini, E., Wyon-Boyault, C., Folco, E., Proponnet-Guerault, M., Vuillerme, N., & Mecella, M. (2026). Quantized medical LLMs for edge deployment: A privacy-preserving RAG system. 2026 IEEE Conference on Artificial Intelligence (CAI), 622–628. <https://doi.org/10.1109/cai68641.2026.11536204>
- [28] Lamaakal, I., Maleh, Y., Makkaoui, K., Ouahbi, I., Plawiak, P., Alfarraj, O., Almousa, M., & El-Latif, A. (2025). Tiny language models for automation and control: Overview, potential applications, and future research directions. *Sensors*, 25(5). <https://doi.org/10.3390/s25051318>