

SARA: An LLM-Based Approach for Systematic Literature Review with a Use Case on Classical–Quantum Integration for Secure Communication

Naiara Santos^{1*}, Marcus Freire^{1,2}, Anderson Tomkelski¹, Thiago Mello^{1,2}, Maycon Peixoto², Ricardo Parizotto¹, João Souza¹

¹QuIIN – Quantum Industrial Innovation, EMBRAP II CIMATEC Competence Center in Quantum Technologies, SENAI CIMATEC University; ²Institute of Computing, Federal University of Bahia (UFBA); Salvador, Bahia, Brazil

This work presents SARA (Semi-Automated Review and Analysis System), an approach that integrates Large Language Models (LLMs), topic modeling, and semantic search to support Systematic Literature Reviews (SLRs). As a use case, SARA was applied to the domain of integration between classical networks and secure quantum communications, covering Quantum Key Distribution (QKD) and Post-Quantum Cryptography (PQC). The system performed deduplication, pre-selection, categorization, and ranking of articles, prioritizing publications with higher potential for scientific contribution. The results indicate a predominance of studies on QKD, followed by PQC and hybrid approaches, with emphasis on telecommunications, IoT, and critical systems. SARA proved effective in identifying publications aligned with the researcher’s queries as well as conceptually related studies, demonstrating its potential to optimize SLR processes in emerging technological areas.

Keywords: Systematic Literature Review. Topic Modeling. Semantic Search. LLMs. Quantum-Safe Communications.

Abbreviations: SARA, Semi-Automated Review and Analysis System. LLM, Large Language Model. QKD, Quantum Key Distribution. PQC, Post-Quantum Cryptography. SLR, Systematic Literature Review. NLP, Natural Language Processing.

The volume of scientific publications has grown exponentially in recent decades, driven by digital dissemination and broad accessibility to specialized databases. This trend makes the execution of Systematic Literature Reviews (SLRs) increasingly challenging, as they demand time, methodological rigor, and substantial human effort. With the advancement of Large Language Models (LLMs), such as GPT and its variants, a growing number of studies explore their use to optimize different phases of the SLR process [1]. Hybrid approaches that combine LLMs with human review have been shown to reduce screening and data extraction time while maintaining high accuracy [2], whereas techniques based on contextual embeddings and deep learning have demonstrated effectiveness in pre-classification and thematic clustering of articles, thereby accelerating the initial screening stages [3].

In recent years, the volume of publications related to quantum technologies applied to information security has grown rapidly, following the maturation of prototypes and the execution of field tests, such as the example of the Rio Quantum Metropolitan Network [4]. This progress is driven both by the race to enable unbreakable communications and by concerns about the so-called “quantum apocalypse,” the possibility that sufficiently powerful quantum computers could compromise currently deployed cryptographic algorithms [5]. Among the approaches to mitigate this risk, two stand out: (i) Quantum Key Distribution (QKD), which uses a quantum channel to securely share cryptographic keys between two parties, leveraging fundamental principles of quantum mechanics and relying on a classical channel for authentication [6]; and (ii) Post-Quantum Cryptography (PQC), which employs encryption algorithms based on mathematical problems considered hard even for quantum computers [7]. Although distinct in their implementation, QKD and PQC pursue the same goal: ensuring that communication remains secure against adversaries equipped with quantum capabilities.

Received on 22 March 2026; revised 15 May 2026.

Address for correspondence: Naiara Santos. Av. Orlando Gomes, 1845 - Piatã, Salvador – BA – Brazil, Zipcode: 41650-010. E-mail: naiara.bonfim@fbter.org.br.

J Bioeng. Tech. Health 2026;9(6):631-637
© 2026 by SENAI CIMATEC University. All rights reserved.

In this context, the need for tools capable of performing pre-selection, classification, and ranking of publications based on semantic criteria becomes evident. This study proposes and applies a semi-automated approach to SLR, which we designate as SARA (Semi-Automated Review and Analysis System). SARA enables the initial screening and categorization of relevant scientific articles by integrating Natural Language Processing (NLP) techniques with LLMs. For this use case, our focus lies on three main objectives: (i) the integration of classical systems with quantum technologies, including PQC, QKD, or both; (ii) the modeling of major recurring topics in the field; and (iii) the semantic retrieval of the most relevant articles, prioritizing those with the greatest potential for scientific contribution.

Related Work

Research on computational methods to support SLRs has evolved rapidly, following the progress of NLP techniques, deep learning, and, more recently, LLMs [8]. Existing approaches vary in scope and degree of automation, ranging from workflows based on traditional machine learning to interactive systems with semantic understanding capabilities. Despite significant advances, few studies fully integrate researcher-driven question-oriented ranking mechanisms, topic modeling, and contextual semantic search—key elements for prioritizing the most relevant publications in complex domains.

As shown in Table 1, Pham and colleagues [9] present a semi-automated workflow combining Random Forest, SVD, LDA, and embeddings, reducing the screening workload by 55–63%. Tsang and Lee [10] combine BERT embeddings, BERTopic, and semantic retrieval, enabling interactive query-based filtering. Ye and colleagues [2] apply bibliometric analysis and topic modeling to map research trends, but without embedding-based semantic retrieval. Ramesh and Sanampudi [11] review active learning for screening, with human feedback influencing prioritization.

Finally, Galli and colleagues [8] integrate LLMs for summarization, classification, and question-guided ranking, combining semantic search with topic modeling.

The SARA approach distinguishes itself by integrating four key elements: LLMs, question-oriented ranking, topic modeling, and semantic retrieval. Unlike prior works as Galli and colleagues [8], SARA links researcher queries directly to corpus themes, enabling dynamic score adjustment and more accurate prioritization, while streamlining pre-selection to reduce human effort. As a toy example for demonstrating SARA, the case study on Classical–Quantum Integration for Secure Communication was based on prior studies on edge–quantum frameworks [12], hybrid cryptographic protocols [13,14], and QKD co-simulation [15]. We present SARA’s architecture, techniques, and implementation, providing a detailed description of SARA’s architecture, the techniques employed, and the implementation process.

Table 1. Comparison of related works.

Reference	Year of publication	Use of LLMs	Question Oriented Ranking	Topic Modeling	Semantic Search
[9]	2021	No	No	Yes	Yes
[10]	2022	No	Partial	Yes	Yes
[11]	2022	No	Partial	No	Partial
[2]	2024	No	No	Yes	Partial
[8]	2025	Yes	Yes	Yes	Yes
SARA	2025	Yes	Yes	Yes	Yes

Material and Methods

The methodology adopted in this study is designed to conduct an SLR in a semi-automated manner, aiming to reduce human effort in the initial stages of article screening and organization.

The methodological workflow comprises four main stages: (i) data collection and deduplication; (ii) LLM-based pre-selection; (iii) topic modeling; and (iv) semantic search (Figure 1). This process, referred to as SARA, spans from the initial deduplication to the final prioritization of articles, incorporating topic modeling and question-guided semantic ranking.

Search String Construction

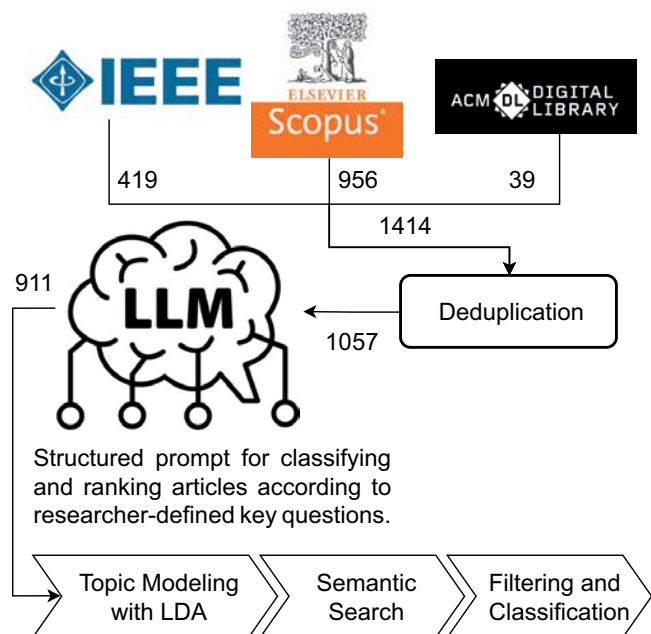
As a preliminary step, a standardized search string was developed to maximize the relevance of results across the three databases consulted (IEEE Xplore, Scopus, and ACM Digital Library),

ensuring the retrieval of publications aligned with the research scope. The query was constructed using Boolean operators applied to the abstract field: ("Quantum Key Distribution" OR "QKD" OR "Quantum Cryptography" OR "Quantum Key Exchange" OR "Post-Quantum Cryptography" OR "PQC" OR "Quantum-Resistant Cryptography" OR "Quantum-Safe Cryptography") AND ("integration" OR "interoperability" OR "migration" OR "framework" OR "compatibility" OR "communication protocol")

The first set of terms refers to the target technologies of this review: QKD, PQC, and terminological variations associated with quantum and post-quantum cryptography; while the second set encompasses concepts related to system integration and interoperability, ensuring that the retrieved results included studies addressing both the technological aspects and their application in communication architectures.

The results obtained from the three databases were exported in CSV format, consolidated, and subjected to a deduplication process based on title, year, and DOI matching (when available, as not all articles provided this identifier), resulting in a total of 1,057 articles.

Figure 1. Methodological workflow of the SARA approach.



The process begins with corpus deduplication, followed by automatic evaluation of the abstracts through a structured prompt, topic modeling with LDA, and semantic search. In the final stage, articles are filtered, classified, and ranked based on the answers to the researcher's questions.

Automated Pre-selection with the LLM (GPT)

A structured prompt was designed for the automatic evaluation of abstracts, covering five criteria: (1) technologies mentioned—PQC, QKD, both, or none; (2) integration with classical networks—real implementation, simulation, conceptual proposal, or no integration; (3) application or prototype—presence and description; (4) technical challenges—quantitative, qualitative, or both; and (5) application domains—up to ten possible areas.

Each criterion was assigned a specific score, with items (1) and (2) defined as exclusion criteria—if both are classified as “None,” the article receives a score of –10 and is automatically excluded from the SLR. Items (3), (4), and (5),

in turn, are classificatory: if the abstract explicitly reports an application, technical challenge, or application domain, 1 point is assigned for each occurrence. Finally, the articles are ranked based on the sum of scores, resulting in a reduced set of 911 studies with potential relevance.

Topic Modeling (Latent Dirichlet Allocation – LDA)

For the thematic organization stage, the LDA algorithm was employed [16], aiming at the automatic detection of predominant topics within the set of pre-selected articles. The abstracts are pre-processed and subsequently submitted to LDA modeling, which enables the identification of latent topics and the clustering of studies into representative macro-themes of the field [17].

Semantic Search Based on Embeddings

After the topic modeling stage with LDA, the results are used to guide the retrieval of semantically relevant documents. For this task, the pre-trained model all-MiniLM-L6-v2, provided by the Sentence-Transformers library, is adopted to convert both abstracts and researcher queries into dense vector representations. The similarity between documents and queries is computed using the cosine metric, enabling the identification of related publications even in the absence of exact lexical matching [18,19].

To enhance retrieval quality, each textual representation was enriched with complementary information automatically extracted by an LLM (ChatGPT-4), incorporating essential aspects such as the application, the type of technology addressed (PQC, QKD, or both), and the proposed level of integration (real, simulated, or conceptual). This semantic enrichment increased the model's ability to capture contextual nuances, resulting in more accurate recommendations aligned with the objectives of the review.

The complete SARA workflow, from the initial deduplication to the final prioritization of articles

is presented in Figure 1, highlighting the stages of topic modeling and question-guided semantic ranking, along with semantic search. The next section presents the results obtained through the application of this workflow, discussing both the performance of individual stages and the overall contributions of the system to conducting Systematic Literature Reviews in the domain of QKD–PQC integration.

Results and Discussion

The application of the SARA approach led to the identification and categorization of a significant set of articles distributed across the technologies considered in this review: QKD, PQC, and hybrid approaches combining both, with the prerequisite of integration into classical networks, that is, the infrastructure and technologies currently in use.

As shown in Figure 2, a predominance of publications exclusively related to QKD was observed (497 articles), followed by PQC (306 articles), and, to a lesser extent, studies addressing QKD and PQC simultaneously (71 articles). This distribution highlights not only the more advanced stage of the literature on QKD but also the growing, albeit still limited, exploration of hybrid solutions integrating both technologies—an aspect particularly relevant to this research, given its focus on secure integration strategies between classical and quantum systems.

The analysis of application domains (Figure 3) shows a predominance of the telecommunications and 5G sector, reflecting the relevance of QKD and PQC for securing high-capacity and low-latency channels. Next, IoT/Edge, government/defense, data centers/cloud, and satellite-based systems stand out, evidencing the interest in strengthening the security of critical networks and sensitive infrastructures. Areas such as blockchain/DLT, industrial systems, the financial sector, and connected vehicles also indicate the broad potential for adopting these technologies within the digital ecosystem.

Figure 2. Distribution of the identified articles by technology: QKD, PQC, and hybrid approaches (QKD+PQC).

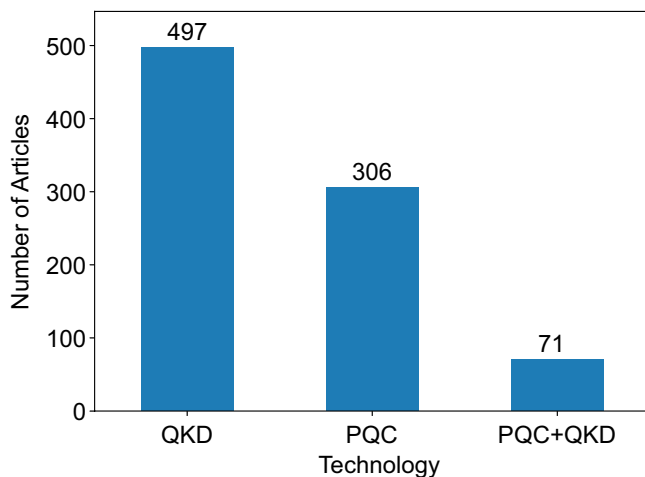
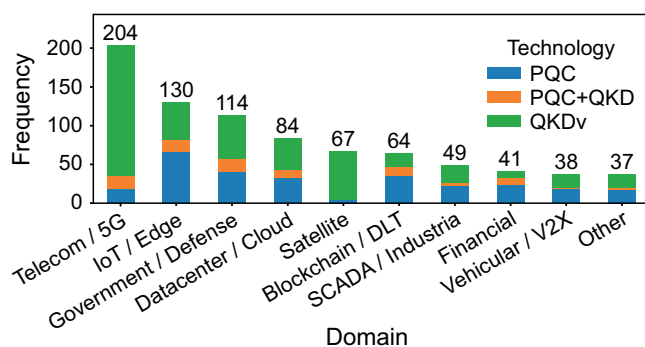


Figure 3. Distribution of articles by application domain of QKD and PQC technologies.



To evaluate the ability of the SARA approach to identify publications aligned with specific researcher queries, the following question was formulated: “What research applies quantum key distribution (QKD) or post-quantum cryptography (PQC) in 5G or telecommunications networks?”. The semantic search, using vector embeddings and cosine similarity calculation, returned a set of articles ranked according to their semantic proximity to the query. Figure 4 shows the two-dimensional projection (PCA) of this vector space, where blue points represent the articles selected as most relevant, and the red point indicates the vector position of the query within the same space. Gray points correspond to non-selected articles. This result demonstrates that SARA not only locates

publications that share terms with the query but also those that exhibit conceptual relationships even with different vocabularies, assigning a quantitative degree of similarity that enables the prioritization of the most pertinent readings.

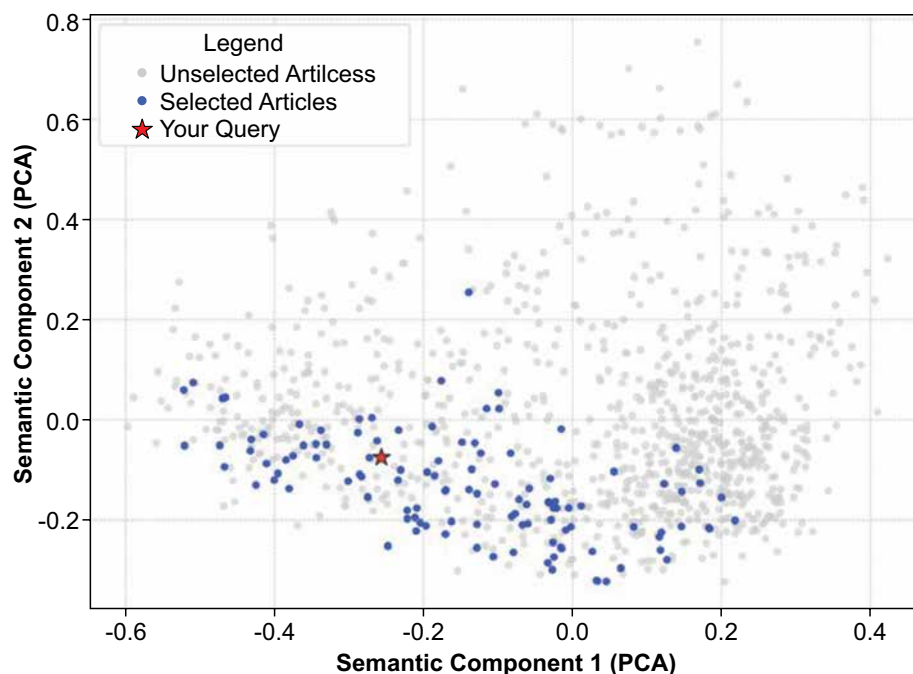
In summary, the results confirm that SARA effectively integrates topic modeling, semantic enrichment, and question-guided ranking, optimizing the selection and prioritization of articles in SLRs. The system demonstrated the ability to retrieve publications aligned with queries and to identify conceptually related studies even in the absence of lexical matching. The predominance of telecommunications and 5G underscores the relevance of QKD and PQC for the security of critical networks, while hybrid solutions emerge as a promising field. These findings consolidate SARA as a useful tool for research in emerging technological domains.

Conclusion

This work presented the SARA approach, a semi-automated methodological workflow to support SLRs, integrating topic modeling, semantic enrichment, and question-oriented ranking, with a focus on the domain of integration between classical networks and secure quantum communications (QKD and PQC). The application of SARA demonstrated that combining NLP techniques with LLMs can significantly reduce human effort in article screening and prioritization while maintaining high relevance in the results. The analyses showed that, in addition to retrieving publications directly aligned with the researcher’s queries, the system is capable of identifying conceptually related studies even in the absence of coinciding terms, an essential aspect in areas with diverse and rapidly evolving technical terminology.

The main contributions are: (i) integrating topic modeling with semantic search for corpus organization; (ii) leveraging LLM-based enrichment to enhance retrieval; and (iii) enabling dynamic question-driven ranking to guide

Figure 4. Two-dimensional projection (PCA) of the vector space resulting from the semantic search for the query “What research applies quantum key distribution (QKD) or post-quantum cryptography (PQC) in 5G or telecommunications networks?”.



Blue points represent the articles selected as most relevant, gray points indicate non-selected articles, and the red star corresponds to the query position in the semantic space.

review focus. The study further underscores the relevance of QKD and PQC for critical sectors, particularly telecommunications, IoT, and hybrid systems.

As future work, we propose: (i) expanding the analyzed dataset by incorporating multiple indexed sources and gray literature; (ii) exploring alternative embedding architectures and retrieval-augmented generation (RAG) models to improve semantic search precision; (iii) incorporating quantitative evaluation metrics such as precision, recall, and F1-score to measure performance objectively; and (iv) adapting SARA to other scientific domains, validating its generalization and comparing it against traditional SLR approaches.

Acknowledgments

This work was partially funded by the QuIN project “Integration of CV-QKD with Classical Networks”, supported by QuIN – Quantum

Industrial Innovation, the EMBRAP II CIMATEC Competence Center in Quantum Technologies, with financial resources from the PPI IoT/Manufacturing 4.0 program under the MCTI call no. 053/2023, established with EMBRAP II. This study was also partially supported by the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES) – Financing Code 001, and by the National Council for Scientific and Technological Development (CNPq), Brazil, under grant no. 403231/2023-0.

References

1. Scherbakov D, Hubig N, Jansari V, Bakumenko A, Lenert LA. The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review. *J Am Med Inform Assoc.* 2025;32(6):1071-1086.
2. Ye A, Maiti A, Schmidt M, Pedersen SJ. A hybrid semi-automated workflow for systematic and literature review processes with large language model analysis. *Future Internet.* 2024;16(5):167.

3. Alchokr R, Borkar M, Thotadarya S, Saake G, Leich T. Supporting systematic literature reviews using deep-learning-based language models. In: *Proceedings of the 1st International Workshop on Natural Language-based Software Engineering*; 2022. p. 67-74.
4. Temporão GP, Melo FRVB, Khoury AZ. The Rio quantum network: a reconfigurable hybrid multiuser metropolitan quantum key distribution network. In: *Workshop de Redes Quânticas*; 2024. p. 19-24.
5. AbuGhanem M. IBM quantum computers: evolution, performance, and future directions. *J Supercomput*. 2025;81(5):687.
6. Qi B. Bennett-Brassard 1984 quantum key distribution using conjugate homodyne detection. *Phys Rev A*. 2021;103(1):012606.
7. Alagic G, Apon D, Cooper D, Dang Q, Dang T, Kelsey J, et al. Status report on the third round of the NIST post-quantum cryptography standardization process. 2022.
8. Galli C, Gavrilova AV, Calciolari E. Large language models in systematic review screening: opportunities, challenges, and methodological considerations. *Information*. 2025;16(5):378.
9. Pham B, Jovanovic J, Bagheri E, Antony J, Ashoor H, Nguyen TT, et al. Text mining to support abstract screening for knowledge syntheses: a semi-automated workflow. *Syst Rev*. 2021;10(1):156.
10. Tsang YP, Lee CKM. Artificial intelligence in industrial design: a semi-automated literature survey. *Eng Appl Artif Intell*. 2022;112:104884.
11. Ramesh D, Sanampudi SK. An automated essay scoring systems: a systematic literature review. *Artif Intell Rev*. 2022;55(3):2495-2527.
12. Maia A, Freire M, Mello T, Rodrigues-Filho R, Almeida E, Prazeres C, et al. Q-Edge: leveraging quantum computing for enhanced software engineering in vehicular networks. In: *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*; 2025. p. 1457-1467.
13. Freire M, Mello TL, Sant'Anna I, Maia A, Moreira R, Rivelino R, et al. RANA: uma abordagem híbrida para QKD BB84 com expansão e encapsulamento de chave. In: *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*; 2025. p. 938-951.
14. Mello TL, Sant'Anna I, Freire M, Maia A, Moreira R, Rivelino R, et al. A new perspective on key expansion for QKD BB84: the RANA model. In: *Workshop de Redes Quânticas (WQuNets)*; 2025. p. 43-48.
15. Tomkelski A, Freire M, Peixoto M, Souza J, Silva V, Parizotto R. Uma abordagem para prototipagem de aplicações com distribuição quântica de chaves. In: *Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais (SBSeg)*; 2025. p. 373-379.
16. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. *J Mach Learn Res*. 2003;3:993-1022.
17. Egger R, Yu J. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. *Front Sociol*. 2022;7:886498.
18. Seo J, Lee S, Liu L, Choi W. TASBERT: Token Attention Sentence-BERT for improving sentence representation. *IEEE Access*. 2022;10:39119-39128.
19. Kamil M, Çakır D. Advances in transformer-based semantic search: techniques, benchmarks, and future directions. *Turk J Math Comput Sci*. 2025;17(1):145-166.