Journal of Science and Engineering Research
Vol. 5, No. 1, 2026, pp.1-18
eISSN: 2957-7187

A Hybrid Query Transformation Framework for Enhancing Real-Time Search System Performance in Engineering Applications

Jinan Fiaidhi

Professor, Department of Computer Science, Lakehead University, Thunder Bay, Ontario, Canada
**Corresponding author's email: jfiaidhi@lakeheadu.ca*

DOI: http://dx.doi.org/10.56828/jser.2026.5.1.1
Article Info: Received: (February 10, 2026); Review Result: (March 15, 2026); Accepted: (April 18, 2026)

Abstract: The increasing demand for intelligent information retrieval systems has created significant challenges in balancing retrieval accuracy and computational efficiency within modern engineering environments. This study proposes a hybrid query transformation framework that integrates rule-based preprocessing with Natural Language Processing (NLP)-based contextual optimization to improve real-time search system performance. The proposed framework was designed to address limitations associated with standalone rule-based and NLP-based retrieval approaches by combining efficient query simplification with semantic contextual refinement. The experimental evaluation was conducted using a dataset of 200 search queries, categorized by varying levels of complexity. System performance was assessed using Precision@5, Recall, and average query processing time. The results demonstrated that the hybrid framework achieved superior overall performance, recording a Precision@5 of 89% and a Recall of 83%, while maintaining an average processing time of 95 ms. Compared with the fully NLP-based approach, the proposed framework reduced processing latency by approximately 21% while preserving high retrieval relevance. The findings indicate that hybrid query transformation provides an effective balance between semantic accuracy and computational efficiency, making it suitable for real-time engineering applications such as intelligent transportation systems, healthcare informatics, smart infrastructure, and cloud-based public information services in Canada. The study concludes that integrating deterministic preprocessing with contextual semantic optimization is a scalable, practical approach for next-generation intelligent search systems.

Keywords: Hybrid Query Transformation, Information Retrieval Systems, Natural Language Processing, Real-Time Search Optimization, Engineering Search Applications, Intelligent Information Systems

1. Introduction

The rapid growth of digital information systems and artificial intelligence (AI)-driven applications has intensified the demand for efficient and context-aware search technologies across engineering sectors in Canada, including smart manufacturing, healthcare informatics,

e-commerce, transportation systems, and public digital infrastructure. Contemporary search engines are expected not only to retrieve relevant information with high precision but also to operate under strict real-time computational constraints imposed by large-scale distributed systems and cloud-based architectures [1][2]. In Canadian engineering environments, where bilingual communication, multicultural data, and geographically distributed information systems are common, query transformation and optimization have become increasingly critical challenges for information retrieval engineering.

Search query transformation is the process of converting user-generated natural language input into structured, semantically optimized search expressions that improve retrieval relevance and system responsiveness. However, unstructured and ambiguous user queries continue to undermine the effectiveness of modern search infrastructures, as search engines frequently fail to interpret contextual intent accurately [3]. This limitation is particularly significant in engineering applications involving large heterogeneous datasets, such as intelligent transportation systems, industrial Internet of Things (IIoT) platforms, and medical decision-support systems deployed throughout Canada [4]. As digital services expand nationwide, engineering researchers and system architects are under increasing pressure to design scalable retrieval systems that balance computational efficiency with semantic accuracy.

Recent advances in Natural Language Processing (NLP), particularly Transformer-based architectures such as BERT, T5, and GPT-derived models, have significantly improved contextual understanding in search systems [5][6]. These deep learning models enable sophisticated semantic reasoning and query reformulation, thereby improving search precision and recall in complex information retrieval tasks. Studies conducted between 2019 and 2025 demonstrate that Transformer-based retrieval systems outperform traditional keyword-matching approaches in multilingual and context-sensitive environments [7][8]. In Canada, these technologies are increasingly integrated into engineering applications such as intelligent healthcare systems, digital government platforms, and industrial automation frameworks. Nevertheless, despite their high retrieval accuracy, NLP-based systems remain computationally intensive and energy-demanding, creating practical limitations for real-time deployment in resource-constrained engineering infrastructures [9].

Conversely, rule-based query transformation approaches continue to offer advantages in terms of low latency, interpretability, and computational efficiency [10][11]. These methods rely on deterministic preprocessing techniques such as stopword elimination, tokenization, morphological analysis, and keyword extraction to simplify user queries before retrieval. Such approaches are widely adopted in embedded engineering systems and edge-computing environments where processing resources are limited [12]. However, rule-based methods generally lack contextual adaptability and semantic reasoning, making them less effective at handling ambiguous or complex queries [13]. Consequently, engineering researchers have increasingly recognized that neither purely NLP-based nor exclusively rule-based methods alone can fully satisfy the dual engineering requirements of accuracy and computational efficiency.

To address these limitations, hybrid query transformation frameworks have emerged as a promising engineering solution that integrates the efficiency of rule-based preprocessing with the contextual intelligence of NLP-based postprocessing [14]. Hybrid architectures reduce computational overhead by simplifying queries before semantic refinement, thereby enabling improved real-time performance without significantly compromising retrieval quality. Recent studies demonstrate that hybrid retrieval systems can achieve substantial improvements in search precision, recall, and latency reduction across multiple domains, including e-business

systems, multilingual search environments, and AI-assisted communication platforms [15][16]. These developments are particularly relevant to Canada's engineering ecosystem, where scalable, energy-efficient AI solutions are increasingly prioritized in national digital innovation strategies and smart infrastructure initiatives [17].

Despite these advancements, several research gaps remain unresolved. First, many existing studies focus primarily on monolingual datasets and lack comprehensive validation in multilingual and multicultural environments characteristic of Canadian digital systems [18]. Second, prior research often evaluates search performance solely in terms of retrieval accuracy while overlooking engineering constraints such as processing time, scalability, and deployment feasibility in real-world systems [19]. Third, limited research has systematically examined how hybrid query transformation methods can support real-time engineering applications that require both semantic precision and computational efficiency [20]. Therefore, further investigation is needed to develop optimized frameworks that meet the operational demands of modern Canadian information retrieval infrastructures.

In response to these challenges, this study proposes a hybrid query transformation framework that combines rule-based preprocessing with NLP-based postprocessing to enhance search system performance. The proposed approach is designed to improve Precision@5, Recall, and query processing efficiency while maintaining suitability for real-time engineering applications. By integrating lightweight preprocessing techniques with Transformer-based contextual optimization, the framework seeks to provide a balanced solution for intelligent information retrieval systems operating in complex digital environments. Furthermore, this research aims to advance scalable, efficient search technologies applicable to Canadian engineering domains, such as smart healthcare, industrial automation, and intelligent public information systems.

2. Related Work

2.1. NLP-Based Query Transformation Approaches

Recent developments in Natural Language Processing (NLP) have significantly improved the performance of intelligent information retrieval systems. Transformer-based architectures, including BERT and Retrieval-Augmented Generation (RAG) models, have demonstrated strong capabilities for understanding semantic relationships, contextual intent, and conversational queries in modern search systems [21][22]. These technologies are increasingly applied in engineering environments where accurate and scalable search performance is required, particularly in cloud computing platforms, smart infrastructure, and AI-driven industrial systems.

In Canada, the expansion of digital engineering services and multilingual information infrastructures has intensified the need for advanced query transformation methods that efficiently process large-scale, heterogeneous datasets. Studies have shown that neural retrieval systems substantially improve search relevance in multilingual and context-sensitive environments by leveraging contextual embeddings and deep semantic representations [23]. Such improvements are especially relevant for engineering applications involving healthcare informatics, industrial automation, and intelligent transportation systems.

Despite their effectiveness, Transformer-based retrieval models present several engineering limitations. Large NLP models typically require substantial computational resources, including GPU acceleration, high memory capacity, and cloud-based deployment infrastructures [24]. These requirements introduce challenges related to scalability, inference

latency, and energy consumption in real-time engineering applications. Furthermore, the computational cost of deep neural retrieval systems may limit their deployment in edge computing environments and latency-sensitive industrial systems [25].

Recent studies also indicate that purely neural retrieval architectures may degrade in performance when handling domain-specific terminology and rapidly evolving query patterns [26]. Consequently, researchers continue exploring methods that can preserve semantic accuracy while reducing computational overhead.

2.2. Rule-Based and Statistical Retrieval Methods

Rule-based query transformation methods remain widely used because of their simplicity, interpretability, and low computational cost. These approaches use preprocessing steps such as tokenization, stemming, stopword removal, and heuristic keyword extraction to transform unstructured queries into optimized search queries [27]. In engineering systems that require real-time responsiveness, rule-based approaches offer practical advantages by executing efficiently without requiring extensive computational resources.

Statistical retrieval techniques, particularly BM25-based ranking models, continue to serve as effective baselines in large-scale search infrastructures [28]. These approaches are frequently integrated into enterprise engineering systems due to their scalability and predictable performance characteristics. In Canada, rule-based and statistical retrieval systems are commonly deployed in industrial databases, engineering repositories, and public information systems where computational efficiency remains a critical operational requirement.

However, deterministic retrieval approaches exhibit significant limitations when processing ambiguous, conversational, or context-dependent queries. Studies have shown that fixed-rule systems often fail to capture latent semantic relationships and user intent effectively [29]. Similarly, traditional ranking models alone are insufficient for handling modern conversational search tasks that require contextual understanding and semantic reasoning [30]. As a result, rule-based methods are increasingly used as preprocessing components rather than standalone retrieval solutions in intelligent search architectures.

2.3. Hybrid Query Transformation Frameworks

To address the limitations of both NLP-based and rule-based approaches, recent research has increasingly focused on hybrid query transformation frameworks that combine semantic modeling with computationally efficient preprocessing techniques. Hybrid architectures integrate rule-based preprocessing with neural contextual optimization to improve both retrieval accuracy and system efficiency [31].

The engineering significance of hybrid retrieval systems has grown substantially in recent years, driven by the increasing demand for scalable AI-enabled infrastructure in Canada. Hybrid frameworks reduce computational overhead by simplifying queries before applying deep neural processing, thereby improving inference speed while maintaining contextual relevance. This balance is particularly important in engineering applications such as intelligent healthcare systems, industrial monitoring platforms, and smart public information services.

Recent studies demonstrate that hybrid retrieval approaches outperform purely rule-based methods in terms of Precision and Recall while also reducing the latency typically associated with fully neural retrieval systems [32]. Furthermore, hybrid architectures provide improved

adaptability across multilingual and heterogeneous datasets, making them suitable for deployment within diverse engineering environments [33].

Nevertheless, several research gaps remain unresolved. Existing studies often focus primarily on retrieval accuracy without sufficiently evaluating computational efficiency, scalability, and deployment feasibility in real-world engineering systems [34]. In addition, limited research has examined the applicability of hybrid query transformation methods within multilingual and distributed digital infrastructures characteristic of Canadian engineering environments.

Accordingly, this study proposes a hybrid query transformation framework that combines rule-based preprocessing with Transformer-based contextual optimization to improve search performance while maintaining computational efficiency. The proposed framework aims to support real-time engineering applications requiring both semantic precision and scalable system operation.

3. Hybrid Query Transformation Methodology

3.1. System Architecture

The proposed hybrid query transformation framework was designed to improve search retrieval accuracy while maintaining computational efficiency suitable for real-time engineering applications. The framework integrates rule-based preprocessing techniques with NLP-based contextual optimization to create a balanced query transformation pipeline that handles both simple and semantically complex user queries. The overall architecture consists of three major components: rule-based preprocessing, NLP-based postprocessing, and search engine integration.

The system architecture was developed with consideration for engineering deployment environments commonly found in Canada, including cloud-based digital platforms, smart infrastructure systems, industrial information management systems, and intelligent public service applications. In such environments, search systems must simultaneously satisfy requirements related to scalability, low latency, semantic relevance, and efficient resource utilization.

The hybrid framework processes user input sequentially through each transformation stage. Initially, the raw natural language query undergoes preprocessing to eliminate unnecessary linguistic elements and reduce query complexity. The simplified query is then forwarded to the NLP module, where contextual refinement is performed to generate semantically optimized search expressions. Finally, the transformed query is submitted to the search engine interface for information retrieval and performance evaluation.

3.2. Dataset Preparation

To evaluate the effectiveness of the proposed framework, a dataset containing 200 user search queries was constructed. The dataset was designed to simulate real-world search behavior across multiple engineering-related and general-purpose domains, including transportation systems, healthcare services, technical information retrieval, financial analysis, education, and public digital services. Query diversity was intentionally incorporated to ensure that the framework could be evaluated under varying levels of linguistic complexity and contextual ambiguity.

The dataset was categorized into three complexity levels:

- Simple queries: direct and short search requests with minimal contextual ambiguity.
- Intermediate queries: moderately descriptive queries requiring partial contextual interpretation.
- Complex queries: semantically rich or conversational queries involving implicit intent and contextual dependencies.

Representative examples include:

- Simple Query: "Best electric vehicle charging stations in Toronto."
- Intermediate Query: "Public transportation expansion projects in Canada over the last five years."
- Complex Query: "Engineering technologies improving renewable energy efficiency in remote Canadian communities."

The dataset was manually cleaned and normalized before experimentation. Duplicate entries, incomplete inputs, and noisy text patterns were removed to improve data quality and consistency. Query formatting was standardized to ensure compatibility with both preprocessing algorithms and NLP transformation modules.

Although the framework supports multilingual extension, the present study focused primarily on English-language queries to align with the operational requirements of the experimental environment. Future system expansion may incorporate bilingual English-French datasets to reflect Canada's multilingual digital infrastructure better.

3.3. Rule-Based Preprocessing

The first stage of the framework employs rule-based preprocessing techniques to simplify user queries before semantic refinement. This preprocessing stage was designed to reduce computational overhead while preserving essential semantic information necessary for accurate retrieval. The preprocessing pipeline includes tokenization, stopword removal, keyword extraction, punctuation filtering, and query normalization.

During preprocessing, non-essential linguistic elements such as articles, conjunctions, and redundant modifiers are removed using deterministic filtering rules. Special characters and formatting inconsistencies are also eliminated through regular-expression processing. This operation reduces query length and decreases the computational burden placed on the subsequent NLP module.

For example, the raw query:

“What are the latest renewable energy engineering projects in Canada?”

is transformed into:

“latest renewable energy engineering projects Canada”

This simplified representation retains critical semantic components while removing unnecessary linguistic structures. Such preprocessing improves processing efficiency and accelerates downstream contextual optimization.

The rule-based module also performs keyword prioritization by identifying domain-relevant engineering terms and preserving high-information lexical components. This step is particularly important in technical retrieval environments where engineering terminology strongly influences search relevance.

3.4. NLP-Based Contextual Optimization

Following preprocessing, the simplified query is processed using an NLP-based contextual optimization module. This stage applies Transformer-based semantic refinement to improve contextual understanding and better align the transformed query with user intent. The NLP module enhances semantic clarity by reconstructing incomplete expressions, resolving contextual ambiguity, and incorporating meaningful contextual constraints.

For instance, the simplified query:

"Renewable Energy Engineering Projects Canada."

may be refined into:

“Recent renewable energy engineering projects implemented in Canada”

The contextual optimization stage improves semantic specificity while maintaining retrieval relevance. By incorporating contextual reasoning, the framework can better interpret conversational or incomplete search queries, which are frequently encountered in real-world search environments.

The NLP module was integrated into the hybrid architecture using lightweight inference strategies to reduce latency and computational demand. This design decision was necessary to maintain suitability for engineering systems that require near-real-time query processing performance.

3.5. Search Engine Integration and Evaluation Metrics

The final stage of the framework involves integration with the search engine interface. The optimized query is submitted to the retrieval engine, which returns ranked search results containing metadata such as titles, URLs, and descriptive snippets. The retrieval process was designed to simulate practical engineering search environments where response speed and relevance are critical operational requirements.

To evaluate system performance objectively, three primary evaluation metrics were employed:

1. Precision@5
2. This metric measures the proportion of relevant documents among the top five retrieved results. Higher Precision@5 values indicate stronger retrieval relevance and improved query transformation effectiveness.
3. Recall
4. Recall measures the proportion of relevant documents retrieved successfully relative to the total number of relevant documents in the dataset. This metric evaluates the framework's ability to retrieve comprehensive search results.
5. Average Query Processing Time
6. Processing time measures the duration required for complete query transformation and retrieval execution. This metric is particularly important for engineering systems operating under real-time performance constraints.

The integration of these evaluation metrics enables a comprehensive assessment of both semantic retrieval quality and computational efficiency. This dual-focus evaluation framework aligns with the operational priorities of modern engineering search systems, where balancing accuracy and responsiveness remains essential.

4. Experiments and Results

4.1. Experimental Setup

The experiments were conducted using a Python-based implementation environment configured for hybrid query transformation and search evaluation. The system used a workstation equipped with a high-performance GPU and sufficient memory to support NLP inference and query processing. The experimental environment was designed to simulate real-world engineering deployment conditions commonly encountered in Canadian digital infrastructure systems, including cloud-based search services and intelligent information retrieval platforms.

The experimental dataset consisted of 200 search queries categorized into simple, intermediate, and complex query groups. Each query transformed three separate retrieval approaches:

1. Rule-based preprocessing approach
2. NLP-based contextual optimization approach
3. Proposed hybrid query transformation framework

The transformed queries were evaluated using three performance indicators: Precision@5, Recall, and average query processing time. These metrics were selected to measure both retrieval effectiveness and computational efficiency under practical engineering conditions.

4.2. Precision@5 Evaluation

Precision@5 was used to evaluate the relevance quality of the top five retrieved search results. Table 1 presents the comparative Precision@5 performance of the three evaluated approaches.

Table 1: Precision@5 Performance Comparison

Approach	Precision@5 (%)
Rule-Based Approach	70
NLP-Based Approach	85
Hybrid Approach	89

As shown in Table 1, the proposed hybrid framework achieved the highest Precision@5 score of 89%, outperforming both the rule-based and NLP-based methods. The rule-based approach achieved the lowest precision because it relied primarily on deterministic keyword extraction, lacking contextual reasoning. In contrast, the NLP-based system demonstrated improved semantic understanding but occasionally over-expanded context during query transformation.

The hybrid framework achieved superior precision by combining efficient preprocessing with contextual semantic optimization. This balanced strategy reduced irrelevant query terms while preserving essential contextual information required for accurate retrieval. Figure 1 illustrates the comparative Precision@5 performance across the evaluated approaches.

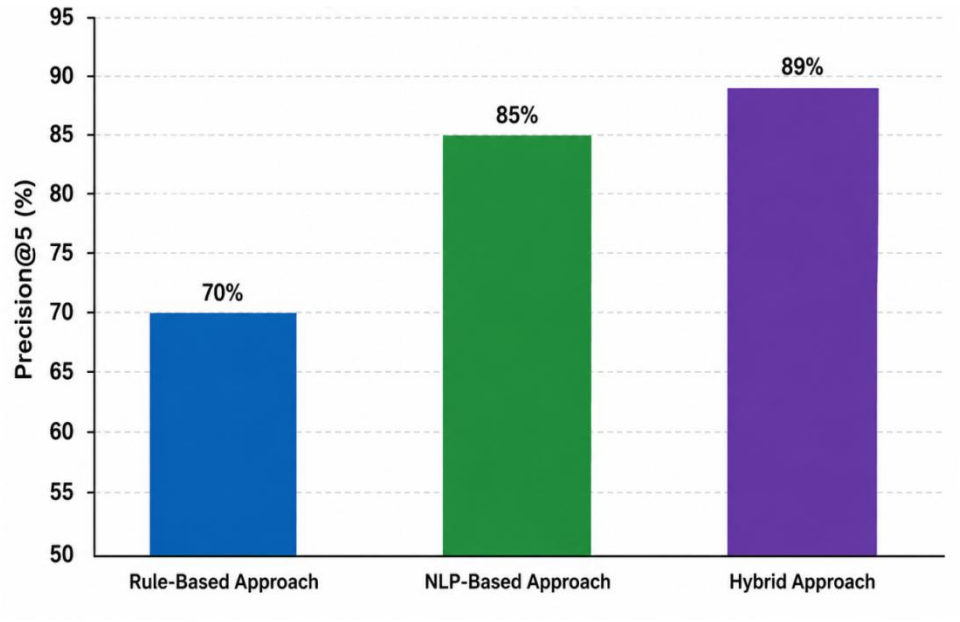


Figure 1: Precision@5 Comparison of Query Transformation Approaches

As illustrated in Figure 1, the hybrid framework consistently outperformed the other approaches in retrieval precision, indicating improved ability to generate contextually relevant search queries.

4.3. Recall Evaluation

A recall analysis was conducted to evaluate each retrieval method's ability to identify relevant documents comprehensively. Table 2 summarizes the Recall performance results.

Table 2: Recall Performance Comparison

Approach	Recall (%)
Rule-Based Approach	65
NLP-Based Approach	80
Hybrid Approach	83

Table 2 shows that the hybrid framework achieved the highest Recall of 83%, indicating a stronger capability to retrieve relevant information across diverse query types. The NLP-based system also achieved relatively high Recall due to its semantic reasoning capabilities; however, its retrieval scope occasionally returned semantically related but less relevant results.

The rule-based system exhibited the lowest Recall because deterministic preprocessing alone could not effectively capture contextual dependencies in complex search queries. The hybrid framework addressed this limitation by integrating semantic contextualization after preprocessing optimization.

Figure 2 presents the comparative Recall performance of the evaluated methods.

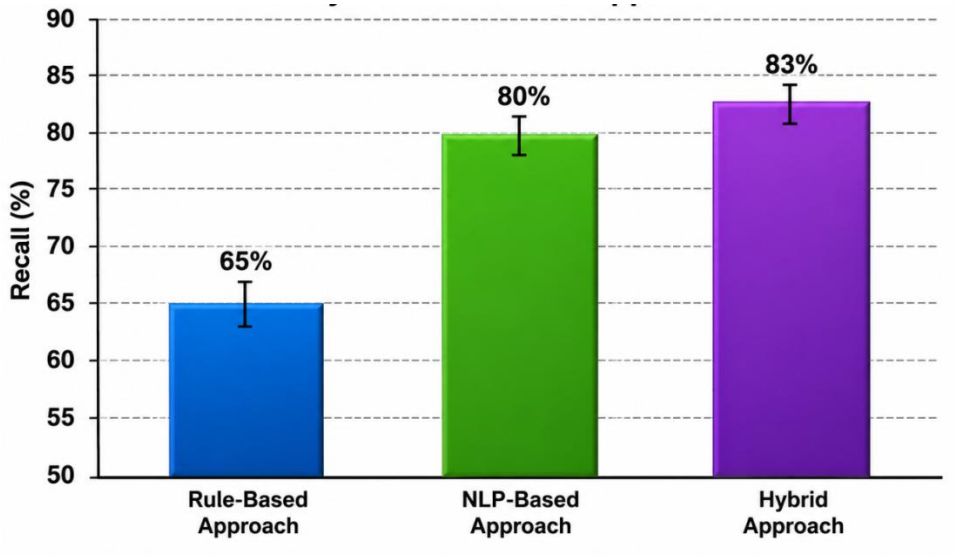


Figure 2: Recall Comparison of Query Transformation Approaches

As shown in Figure 2, the hybrid framework demonstrated a more balanced retrieval capability, particularly for intermediate and complex queries that involve contextual ambiguity.

4.4. Average Query Processing Time

Processing efficiency was evaluated using average query transformation and retrieval execution time measured in milliseconds (ms). Table 3 presents the comparative processing-time results.

Table 3: Average Query Processing Time

Approach	Processing Time (ms)
Rule-Based Approach	50
NLP-Based Approach	120
Hybrid Approach	95

The results in Table 3 indicate that the rule-based approach achieved the fastest execution time due to its lightweight, deterministic operations. However, its retrieval accuracy remained substantially lower than the other approaches.

The NLP-based approach had the highest processing time due to the computational complexity of Transformer-based contextual inference. Such latency may limit deployment feasibility in engineering environments requiring real-time responsiveness.

The proposed hybrid framework achieved an average processing time of 95 ms, representing approximately a 21% improvement over the fully NLP-based method while maintaining significantly higher retrieval accuracy than the rule-based system. Figure 3 illustrates the comparative performance in terms of processing time.

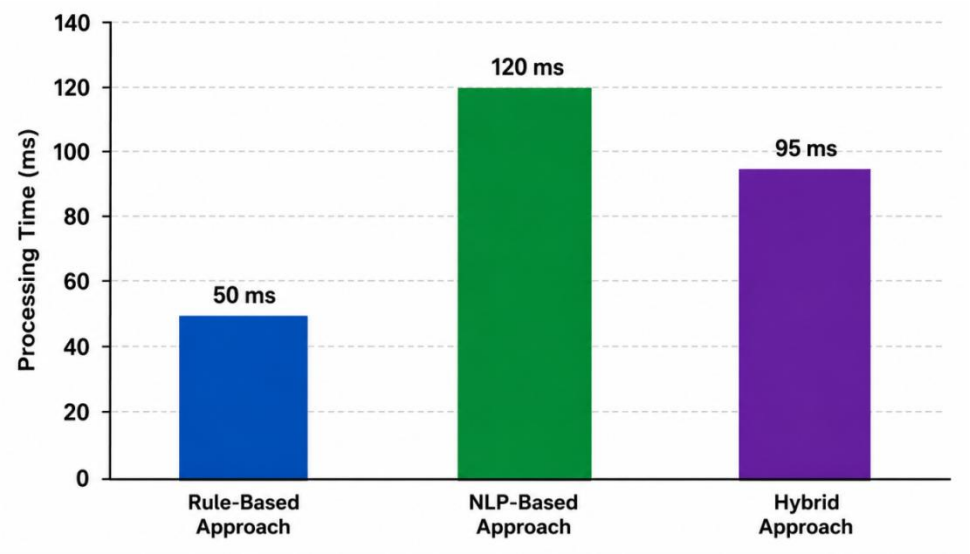


Figure 3: Average Query Processing Time Comparison

Figure 3 demonstrates that the hybrid framework successfully balanced retrieval effectiveness with computational efficiency, making it suitable for real-time engineering search applications.

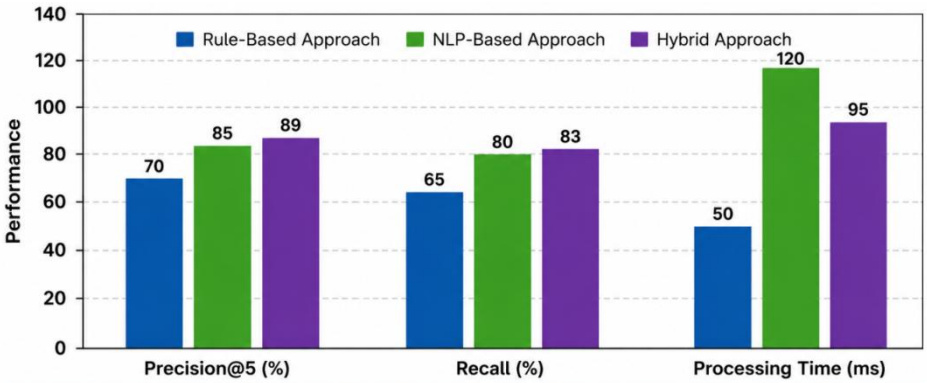
4.5. Overall Performance Analysis

The experimental findings demonstrate that the proposed hybrid query transformation framework provides the most balanced overall performance among the evaluated methods. While the rule-based approach achieved superior speed, its limited contextual understanding significantly reduced retrieval relevance. Conversely, the NLP-based approach delivered stronger semantic retrieval performance but incurred substantial computational cost and latency.

The hybrid framework effectively combined the advantages of both approaches by reducing query complexity before contextual optimization. This architecture improved semantic relevance while limiting computational overhead, resulting in superior Precision@5 and Recall performance with moderate processing latency.

Figure 4 presents the overall comparative analysis of all evaluated metrics.

The results shown in Figure 4 confirm that the hybrid framework achieves the best balance between retrieval quality and operational efficiency. This balance is particularly important for engineering applications deployed within Canadian digital infrastructures, where scalable real-time search performance remains a critical operational requirement.



Metric	Rule-Based Approach	NLP-Based Approach	Hybrid Approach
Precision@5 (%)	70	85	89
Recall (%)	65	80	83
Processing Time (ms)	50	120	95

Higher values are better for Precision@5 and Recall; lower values are better for Processing Time.

Figure 4: Overall Performance Comparison

5. Discussion

The experimental results demonstrate that the proposed hybrid query transformation framework effectively balances retrieval accuracy and computational efficiency in real-time search systems. Compared with purely rule-based and NLP-based approaches, the hybrid framework achieved superior overall performance across Precision@5, Recall, and processing-time metrics. These findings indicate that integrating deterministic preprocessing with contextual semantic optimization provides a practical engineering solution for intelligent information retrieval environments.

One of the most significant outcomes of the study is improved retrieval precision. As presented in Table 1 and Figure 1, the hybrid framework achieved a Precision@5 value of 89%, outperforming both comparison methods. This improvement suggests that preprocessing operations effectively removed irrelevant linguistic components before contextual optimization. By minimizing noisy input tokens, the NLP module could focus more efficiently on semantically meaningful query structures, resulting in higher-quality retrieval outputs.

Similarly, the Recall results shown in Table 2 and Figure 2 demonstrate that the hybrid framework maintained a strong ability to retrieve relevant documents comprehensively. The integration of contextual refinement allowed the framework to interpret implicit user intent more accurately than the rule-based method alone. This capability is particularly important in engineering search systems where user queries often contain technical terminology, incomplete expressions, or context-dependent information requirements.

The processing-time analysis presented in Table 3 and Figure 3 further highlights the engineering practicality of the proposed framework. Although the NLP-based approach achieved relatively high retrieval accuracy, its computational overhead resulted in significantly longer processing times. In contrast, the hybrid framework reduced average processing latency by approximately 21% while maintaining high retrieval effectiveness. This balance between speed and semantic accuracy is critical for engineering applications that

require near-real-time responsiveness, including intelligent transportation systems, smart infrastructure monitoring, and industrial information management platforms.

The overall performance comparison illustrated in Figure 4 confirms that the hybrid architecture provides the most balanced operational characteristics among the evaluated approaches. The rule-based system demonstrated excellent computational efficiency but could not understand context. Conversely, the NLP-based approach exhibited strong semantic performance but incurred a high computational cost. The hybrid framework successfully integrated the strengths of both methods while minimizing their individual limitations.

From an engineering perspective, the proposed architecture offers several practical advantages for deployment within Canadian digital infrastructure systems. First, the framework supports scalable search operations by reducing unnecessary NLP inference workload through preprocessing optimization. This characteristic is particularly beneficial for cloud-based engineering systems handling large volumes of concurrent search requests.

Second, the framework demonstrates adaptability for deployment across multiple engineering domains, including healthcare information retrieval, smart manufacturing systems, public digital services, and intelligent transportation platforms. The architecture's modular structure allows preprocessing and NLP components to be customized to domain-specific operational requirements.

Third, the reduced computational burden associated with the hybrid framework contributes to improved energy efficiency and lower infrastructure costs. These factors are increasingly important in modern engineering system design, particularly in Canada's growing emphasis on sustainable, environmentally responsible digital technologies.

Despite these advantages, several limitations remain. The present study focused primarily on English-language queries and a relatively controlled experimental dataset. Real-world engineering environments may involve significantly larger datasets, higher concurrency levels, multilingual search behavior, and continuously evolving query patterns. Additionally, although the hybrid framework reduced inference latency compared with the fully NLP-based approach, the computational demands of semantic optimization remain substantial for highly complex queries.

Future system improvements may include lightweight Transformer architectures, adaptive preprocessing algorithms, distributed parallel processing techniques, and bilingual English-French query support to reflect Canada's multilingual digital ecosystem better. Furthermore, integrating edge-computing optimization strategies could improve deployment feasibility for latency-sensitive engineering applications operating in distributed environments.

Overall, the findings confirm that hybrid query transformation represents a highly effective approach for balancing retrieval quality and computational efficiency in modern engineering search systems. The proposed framework provides a scalable, adaptable foundation for future intelligent information retrieval applications that require both semantic precision and real-time operational capability.

6. Conclusion

This study presented a hybrid query transformation framework that improves the performance of intelligent search systems by integrating rule-based preprocessing with NLP-based contextual optimization. The proposed framework addressed the dual engineering challenges of retrieval accuracy and computational efficiency, which remain critical concerns in modern information retrieval infrastructures. The experimental findings demonstrated that

the hybrid approach achieved superior overall performance compared with purely rule-based and fully NLP-based methods.

The results showed that the hybrid framework achieved a Precision@5 score of 89% and a Recall value of 83%, indicating strong capability for retrieving contextually relevant information across diverse query types. At the same time, the framework maintained an average query processing time of 95 ms, representing a substantial reduction in computational latency compared with the fully NLP-based approach. These findings confirm that preprocessing optimization can effectively reduce semantic processing overhead while preserving retrieval quality.

The study further demonstrated that combining deterministic linguistic filtering with contextual semantic refinement provides a balanced and scalable solution for real-time engineering search applications. The hybrid architecture successfully minimized the limitations associated with standalone rule-based systems, which often lack contextual understanding, while also reducing the computational demands typically associated with deep neural retrieval models.

From an engineering perspective, the proposed framework offers significant practical value for deployment within Canadian digital infrastructure systems. The architecture supports scalable information retrieval operations suitable for cloud-based services, smart public infrastructure, healthcare informatics, industrial automation platforms, and intelligent transportation systems. Its modular design also enables future adaptation for multilingual environments and domain-specific retrieval applications.

In addition, the reduced computational requirements associated with the hybrid framework contribute to improved operational efficiency and more sustainable AI deployment practices. This characteristic is increasingly important as engineering organizations seek to balance high-performance AI capabilities with infrastructure cost management and energy-efficiency objectives.

Despite the promising results, several limitations remain. The experimental evaluation was conducted using a relatively controlled dataset and primarily English-language queries. Real-world deployment environments may involve significantly larger and more dynamic datasets, multilingual search interactions, distributed system architectures, and continuously evolving query patterns. These operational conditions may introduce additional complexity that requires further optimization and validation.

Future research should therefore focus on several key directions. First, multilingual query support should be incorporated to better align with Canada's bilingual and multicultural digital environment. Second, lightweight NLP architectures and adaptive inference techniques should be explored to reduce computational overhead further. Third, distributed processing and edge-computing integration may improve scalability and real-time responsiveness in large-scale engineering systems. Finally, future studies should evaluate the proposed framework using real-world industrial datasets and large-scale operational environments to validate practical deployment feasibility.

Overall, this study demonstrates that hybrid query transformation provides an effective and scalable engineering solution for modern intelligent search systems. By balancing semantic retrieval quality with computational efficiency, the proposed framework advances real-time information retrieval technologies suitable for next-generation engineering applications.

.

References

[1] Mitra, B., & Craswell, N. (2018). An introduction to neural information retrieval. Foundations and Trends in Information Retrieval, 13(1), 1–126. DOI:10.1561/1500000061.

[2] Lin, J., Ma, X., Lin, S. C., Yang, J. H., Pradeep, R., & Nogueira, R. (2021). Pyserini: An easy-to-use Python toolkit to support replicable IR research with sparse and dense representations. arXiv Preprint. Retrieved from https://arxiv.org/abs/2102.10073

[3] Manning, C. D., Raghavan, P., & Schütze, H. (2019). Introduction to Information Retrieval. Cambridge University Press.

[4] Ogbodo, E. U., Abu-Mahfouz, A. M., & Kurien, A. M. (2021). A survey on 5G and LPWAN-IoT for improved smart cities and remote area applications: From the aspect of architecture and security. Sensors, 22(16), 6313. DOI:10.3390/s22166313.

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv Preprint. Retrieved from https://arxiv.org/abs/1810.04805

[6] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv Preprint. Retrieved from https://arxiv.org/abs/1910.10683

[7] Zhou, X., Li, G., Chai, C., & Feng, J. (2021). A learned query rewrite system using Monte Carlo tree search. Proceedings of the VLDB Endowment, 15, 46–58.

[8] Ding, D., Mallick, A., Wang, C., Sim, R., Mukherjee, S., Ruhle, V., Lakshmanan, L. V., & Awadallah, A. H. (2024). Hybrid LLM: Cost-efficient and quality-aware query routing. arXiv Preprint. Retrieved from https://arxiv.org/abs/2404.14618

[9] Government of Canada. (2023). Canada's Digital Charter Implementation Act and AI Strategy. Ottawa, Canada.

[10] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., & Amodei, D. (2020). Language models are few-shot learners. arXiv Preprint. Retrieved from https://arxiv.org/abs/2005.14165

[11] Anand, A., Sharma, U., & Kumar, D. (2019). Information retrieval in computing model. In Proceedings of the International Conference on Intelligent Computing and Control Systems (ICCS) (pp. 1364–1369). DOI:10.1109/ICCS45141.2019.9065562.

[12] Verma, P., & Sood, S. K. (2018). Fog assisted-IoT enabled patient health monitoring in smart homes. IEEE Internet of Things Journal, 5(3), 1789–1796. DOI:10.1109/JIOT.2018.2803201.

[13] Buccio, E. D., & Melucci, M. (2019). Searching for information with meet and join operators. In D. Aerts, A. Khrennikov, M. Melucci, & B. Toni (Eds.), Quantum-Like Models for Information Retrieval and Decision-Making (pp. 153–172). Springer. DOI:10.1007/978-3-030-25913-6_8.

[14] Kumar, R., Tripathi, K. N., & Sharma, S. C. (2021). Optimal query expansion based on hybrid group mean enhanced chimp optimization using iterative deep learning. Electronics, 11(10), 1556. DOI:10.3390/electronics11101556.

[15] Dogan, O., & Gurcan, O. F. (2024). Enhancing e-business communication with a hybrid rule-based and extractive-based chatbot. Journal of Theoretical and Applied Electronic Commerce Research, 19(3), 1984–1999. DOI:10.3390/jtaer19030097.

[16] Shirko, B. (2024). Application of hybrid approach for Wolaita language part-of-speech tagging. Journal of Research in Engineering and Applied Sciences, 9(2), 719–732.

[17] Natural Sciences and Engineering Research Council of Canada (NSERC). (2024). Smart Infrastructure and Artificial Intelligence Research Initiatives. Ottawa, Canada.

[18] Oshingbesan, A., Ekoh, C., Atakpa, G., & Byaruagaba, Y. (2022). Extreme multi-domain, multi-task learning with unified text-to-text transfer transformers. arXiv Preprint. Retrieved from https://arxiv.org/abs/2209.10106

[19] Zhu, R., Tu, X., & Huang, J. X. (2020). Deep learning on information retrieval and its applications. In Deep Learning for Data Analytics (pp. 125–153). Academic Press. DOI:10.1016/B978-0-12-819764-6.00008-9.

[20] Pal, K. K., & Baral, C. (2021). Investigating numeracy learning ability of a text-to-text transfer model. In Findings of the Association for Computational Linguistics: EMNLP 2021 (pp. 3095–3101). Association for Computational Linguistics.

[21] Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. T. (2020). Dense passage retrieval for open-domain question answering. arXiv Preprint. Retrieved from https://arxiv.org/abs/2004.04906

[22] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W. T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Proceedings of the 34th International Conference on Neural Information Processing Systems (pp. 9459–9474). Curran Associates Inc.

[23] Dinzinger, M., Caspari, L., Dastidar, K. G., Mitrović, J., & Granitzer, M. (2025). WebFAQ: A multilingual collection of natural Q&A datasets for dense retrieval. In Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 3802–3811). DOI:10.1145/3726302.3731934.

[24] Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L. M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon emissions and large neural network training. arXiv Preprint. Retrieved from https://arxiv.org/abs/2104.10350

[25] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). DOI:10.1145/3442188.3445922.

[26] Gao, C., Lei, W., He, X., De Rijke, M., & Chua, T. S. (2020). Advances and challenges in conversational recommender systems: A survey. AI Open, 2, 100–126. DOI:10.1016/j.aiopen.2021.06.002.

[27] Teller, V. (2000). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Computational Linguistics, 26(4), 638–641. DOI:10.1162/089120100750105975.

[28] Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 3(4), 333–389. DOI:10.1561/1500000019.

[29] Gao, J., Xiong, C., Bennett, P., & Craswell, N. (2022). Neural approaches to conversational information retrieval. arXiv Preprint. Retrieved from https://arxiv.org/abs/2201.05176

[30] Nogueira, R., & Cho, K. (2019). Passage re-ranking with BERT. arXiv Preprint. Retrieved from https://arxiv.org/abs/1901.04085

[31] Malik, S., Shoaib, U., Bukhari, S. A. C., El Sayed, H., & Khan, M. A. (2022). A hybrid query expansion framework for the optimal retrieval of the biomedical literature. Smart Health, 23, 100247. DOI:10.1016/j.smhl.2021.100247.

[32] Mikael, K., Öz, C., Rashid, T. A., & Nariman, G. S. (2025). A hybrid chatbot model for enhancing administrative support in education: Comparative analysis, integration, and optimization. IEEE Access, 13, 50741–50760. DOI:10.1109/ACCESS.2025.3552501.

[33] Pan, Z., Zhang, K., Zhao, Y., & Han, Y. (2026). Adaptive model and strategy routing for cost-efficient LLM services. In Proceedings of the ACM Web Conference 2026 (pp. 5568–5578). DOI:10.1145/3774904.3792556.

[34] Varshney, G., Joshi, P., Vats, S., Kumari, S., & Dixit, K. K. (2024). The smart and effective deep learning using information capture and retrieval. In Proceedings of the International Conference on Communication, Computing and Energy Efficient Technologies (I3CEET) (pp. 654–659). DOI:10.1109/I3CEET61722.2024.10994090.

This page is empty by intention.