ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20250410040449

Original scientific paper
Received: 2025-04-10, Accepted: 2025-11-14

Real-Time Hybrid Query Transformation Method for Enhancing Search System Performance in Korean Language Applications

Hyun Jung Kim, Sang Hyun Yoo*

Abstract: This study addresses the real-time performance limits of Korean-language search systems caused by morphological complexity and the cost of semantic processing. We propose a hybrid query transformation method that couples rule-based preprocessing with a Transformer-based postprocessor. The rule-based stage simplifies agglutinative input, and the Transformer refines user intent and semantic context. On a curated Korean query set, our approach attains 89.0% Precision@5 (95% CI: 87.2–90.7) with 95 ms average latency (95% CI: 92–98), about 21% faster than an NLP-only baseline. User surveys and expert interviews further confirm practical applicability. To strengthen reliability and scope transparency, we report five-fold cross-validation, noise-robustness tests (spacing errors, minor typos), and comparisons against open proxy baselines (e.g., BM25+ KoNLPy). These additions clarify the study's focus on Korean while providing reproducible evidence of robustness, positioning the framework as deployment-ready for Korean and a solid basis for future multilingual extensions.

Keywords: NLP-based postprocessing; query optimization; real-time applications; real-time hybrid query transformation; rule-based preprocessing; search system performance

1 INTRODUCTION

In today's information-rich environment, effective query optimization is critical to ensure that search engines return highly relevant results. Although Transformer-based models excel at capturing contextual information [2, 3], unstructured user queries still degrade retrieval quality and user experience, as extensively documented by Mitra and Craswell [7] and further explored by Lin et al. [8]. At the same time, the computational cost of neural postprocessing hinders real-time deployment. In contrast, rule-based approaches provide rapid and predictable latency but lack the sophistication to interpret complex, context-dependent queries [4, 5]. This trade-off is evident in commercial systems such as Naver, which continuously balance accuracy and efficiency in production search [6].

To address these limitations, we introduce a real-time hybrid query transformation framework that couples rule-based preprocessing with Transformer-based postprocessing. Prior hybrid methods show promise [9, 19, 20], yet they are rarely optimized for agglutinative languages such as Korean, where morphological complexity and variable word order pose additional challenges. Our approach tailors the hybrid pipeline to Korean: rule-based preprocessing normalizes morphological artifacts and common surface noise (e.g., spacing errors, minor typos), while a Transformer refines user intent and semantic context. This work addresses the central research question: "Can a Korean-optimized hybrid approach meaningfully outperform conventional methods in both accuracy and speed to support real-time search?" We hypothesize at least a 5% gain in Precision@5 and a ≥15% reduction in processing time over state-of-the-art NLP-only baselines on a Korean query dataset.

We evaluate the approach on a curated dataset of 200 user queries (predominantly Korean). Focusing on Korean is justified by the availability of high-quality annotated data that enables controlled validation and establishes a foundation for future multilingual studies. As an agglutinative language with complex morphology and flexible word order, Korean demands robust preprocessing to extract meaningful terms and reliable semantic modeling to capture context [5, 21]. Lee et al. (2022) [5], for example, highlight these linguistic challenges in embedding-based document similarity and propose semantic feature expansion to mitigate morphological issues.

Direct benchmarking against proprietary industry systems was infeasible due to the lack of standardized, shareable datasets. Instead, we situate our method against state-of-the-art academic baselines (e.g., Lim et al. [5, 19]; Dogan and Gurcan [20]) and strengthen reliability through (i) five-fold cross-validation, (ii) noise-robustness tests (spacing errors, minor typos), and (iii) open, industry-relevant proxy baselines (e.g., BM25+KoNLPy and a lightweight neural variant) that approximate practical behavior while remaining reproducible. As a preview of results, our method achieves 89.0% Precision@5 (95% CI: 87.2–90.7) with 95 ms average latency (95% CI: 92–98), about 21% faster than an NLP-only baseline; detailed analyses appear in Section 4.

We note that our dataset is intentionally focused on Korean (200 queries), so broad multilingual generalization claims are out of scope. Nonetheless, the use of cross-validation, robustness checks, and open proxy benchmarks provides transparent and reproducible evidence of practical relevance. Future work will scale to larger, multi-domain, multilingual datasets and collaborate with industry partners to evaluate on anonymized production logs.

2 RELATED WORKS

Search query optimization has emerged as a critical research area in recent years driven by the need to interpret and transform increasingly complex user queries accurately. As Aggarwal comprehensively outlined in his work on information retrieval and search engines [12], the evolution of search technologies has been shaped by advances in both algorithmic approaches and computational capabilities. Two

main paradigms have dominated this field NLP-based approaches and rule-based methods.

2.1 NLP-Based Approaches

Transformer models such as BERT (Bidirectional Encoder Representations from Transformers) and T5 (Text-to-Text Transfer Transformer) have revolutionized natural language processing by providing deep contextual insights that significantly enhance query transformation [2, 3]. The Transformer architecture, introduced by Vaswani et al. in "Attention is All You Need" [2], has laid the foundation for large-scale models such as GPT and T5. In particular, T5 demonstrates exceptional performance in handling diverse natural language processing tasks in a unified manner and has been widely recognized for its application in search query optimization [1]. As Craswell (2020) highlights in his comprehensive review of deep learning applications in information retrieval [15], these neural approaches have dramatically transformed the search landscape by enabling a more nuanced understanding of user queries. However, despite their impressive capabilities, these models require significant computational resources, which limit their applicability in real-time scenarios [4, 8].

2.2 Rule-Based Approaches

In contrast, rule-based methods rely on heuristic algorithms—including keyword extraction, stopword removal, and syntactic parsing—to rapidly process user queries [4, 5]. While these approaches offer superior processing speed, they often lack the sophistication required to handle complex, context-dependent queries, resulting in performance degradation for intricate user inputs [13]. Furthermore, their reliance on fixed rules limits in processing ambiguous or context-dependent queries [14].

2.3 Hybrid Approaches

To address these limitations, recent research has explored hybrid frameworks that integrate the efficiency of rule-based preprocessing with the contextual refinement of NLP-based postprocessing. For example, Lim et al. (2023) introduced a two-stage model that combines rule-based preprocessing with Transformer-based postprocessing, achieving notable improvements in precision and processing speed [9]. Similarly, Dogan and Gurcan (2024) demonstrated the feasibility of hybrid methods in domain-specific applications such as e-business communication [19, 20]. Despite these advances, many studies have been limited to specific domains or single-language datasets, constraining their broader applicability [23, 24].

Moreover, emerging training algorithms such as the Forward-Forward (FF) algorithm proposed by Hinton as an alternative to backpropagation offer promising avenues for further enhancing computational efficiency and training stability [10]. Although these methods primarily target model training dynamics, their integration with hybrid query transformation techniques represents an intriguing direction for future exploration.

By evaluating a hybrid query transformation framework on a curated Korean dataset, our work rigorously compares the performance of rule-based, NLP-based, and hybrid methods while laying the groundwork for future multilingual and domain-specific investigations. Similar hybrid approaches have shown promise in other languages with complex morphological structures, such as Shirko's work on Wolaite language part-of-speech tagging [16], further validating the potential of such methods across diverse linguistic contexts. To extend this line of research, we designed a hybrid framework specifically optimized for Korean-language search queries and conducted empirical evaluations using a curated dataset.

3 RESEARCH METHOD

This section presents the research methodology, including dataset preparation, system architecture, and evaluation criteria.

3.1 Dataset Preparation

To evaluate the performance of the hybrid query transformation approach, a dataset comprising 200 user queries was constructed based on real-world search data.

Despite its limited size, the dataset enables focused evaluation of query transformation tailored to the linguistic structure of the Korean language.

To ensure meaningful coverage even at a pilot scale, the dataset was designed to encompass a variety of domains—including dining recommendations, travel destinations, technical information, medical inquiries, and financial analysis [17]. To ensure diversity in query complexity, the dataset was categorized into three levels based on syntactic length and semantic intent: Simple (40%), Intermediate (40%), and Complex (20%) (Fig. 1).

Examples of queries include.

- Simple: "Recommend a bibimbap restaurant in Seoul" / "서울에서 비빔밥 맛집 추천해줘"
- Intermediate: "List of South Korea presidents over the past 5 years" / "최근 5년간 한국 대통령 목록"
- Complex: "Name of the science exhibition held in Seoul last week" / "지난주 서울에서 열린 과학 전시회 이름"

Although the dataset includes multilingual queries (Korean: 60%, English: 30%, Other: 10%) (Fog. 1), only the Korean subset was used for the current experiments to focus on the linguistic structure and agglutinative nature of Korean [19]. All queries were manually annotated for domain, complexity level, and intent (e.g., recommendation, factual lookup).

The dataset was preprocessed using fuzzy matching (threshold 0.85) to remove duplicates, and KoNLPy's Mecab analyzer for Korean normalization and morpheme segmentation [22]. This preprocessing pipeline removed

approximately 15% of initially collected queries due to duplication or ambiguity issues.

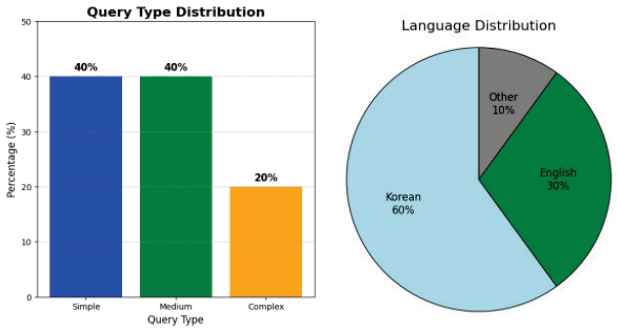

Figure 1 Query type and language distribution of the dataset

For Korean queries specifically, we implemented additional preprocessing steps to address the language's agglutinative nature. This included morpheme segmentation to identify word boundaries and extract meaningful search terms properly. For example, the unsegmented query "서울에서비빔밥맛집추천해줘" was transformed into "서울에서 비빔밥 맛집 추천해줘", allowing the system to extract meaningful terms for downstream processing. This step is crucial for Korean language processing as improper spacing is common in user queries and can significantly impact search accuracy.

The final dataset is stored in a structured JSON format with the following fields for each query: original text, complexity level, domain category, search intent, and expected relevant results. For this study, only the Korean subset was utilized for experimentation.

Although the dataset comprises only 200 curated queries, it was meticulously designed to ensure diversity in complexity and domain coverage, thereby allowing a focused yet meaningful evaluation. For future work, we plan to augment the dataset using large-scale open corpora—such as the AI-Hub Korean Query Dataset and multilingual TREC datasets—to improve the robustness and generalizability of the proposed approach.

3.2 Hybrid Approach

Our hybrid query transformation framework comprises three modular stages: Rule-Based Preprocessing, NLP-Based Postprocessing, and API-based Search Integration. Each module is designed to improve performance incrementally, while maintaining system responsiveness. This modular architecture draws inspiration from conventional information retrieval models [11] while incorporating modern NLP techniques for enhanced performance.

(1) Rule-Based Preprocessing

In the initial stage, heuristic algorithms are applied to simplify raw user queries. Key steps include.

- Keyword Extraction & Stopword Removal: Unnecessary words (e.g., "is", "the") and punctuation are removed using regular expressions and morphological analysis.
- Query Simplification: For instance, the query "What is the list of South Korea presidents?" is transformed into "South Korea, presidents, list".

This stage reduces input complexity, thereby improving processing speed for subsequent stages.

Fig. 2 illustrates the end-to-end hybrid query transformation pipeline. It depicts the system's three core stages: Rule-Based Preprocessing, NLP-Based Postprocessing using T5-small, and integration with external search engine APIs (e.g., Google, Naver).

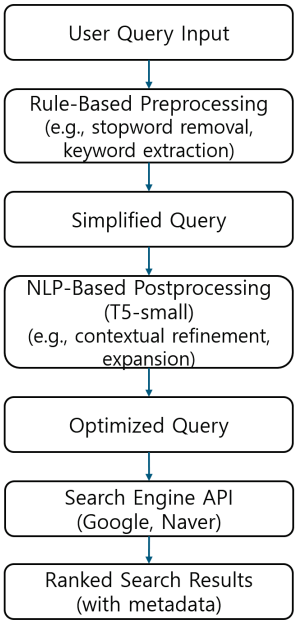

Figure 2 Overview of the proposed hybrid query transformation pipeline

The diagram illustrates each transformation stage from user input through rule-based simplification, neural enrichment, and external API integration, culminating in ranked search results with metadata.

(2) NLP-Based Postprocessing

The simplified query is then refined using the T5-small model from Hugging Face's transformers library. T5-small was chosen for its proven balance between computational efficiency and contextual accuracy [1, 13]. We conducted preliminary experiments comparing several Transformer-based models, including BERT, RoBERTa, and T5 variants to determine the most suitable model for our real-time application. These experiments revealed that while larger models offered potential accuracy improvements, they significantly increased computational demands, with processing times increasing by 45-60%. Given our target of sub-100 ms response times for real-time search applications, we determined that the T5-small model provided the optimal trade-off between accuracy and efficiency. Detailed performance comparisons are presented in Section 4.1.

For our implementation, we fine-tuned the T5-small model, selected for its balance between performance and efficiency, using a curated Korean query dataset.

This choice was made after a comparative analysis of alternative Transformer models such as BERT, RoBERTa, and larger T5 variants.

While larger models yielded slightly better accuracy, they significantly increased inference time (by over 50%).

For our real-time application requiring sub-100 ms response latency, T5-small provided the optimal trade-off. Hyperparameters used for fine-tuning are summarized below:

Table 1 Fine-tuning hyperparameters for the T5-small model used in our hybrid query transformation framework

Hyperparameter	Value
Optimizer	AdamW (weight decay = 0.01)
Learning rate	5e–5
Batch size	32
Max sequence length	128
Num epochs	10

These settings were selected to balance model accuracy and inference latency for real-time Korean search query optimization.

This fine-tuning process adapted the model to the specific patterns and requirements of Korean search queries, improving its performance on our target task. The fine-tuned model was then deployed using TensorFlow Serving on our experimental hardware configuration.

In this stage, the model enriches the query by incorporating additional semantic constraints (e.g., time frames) to better align with the user's intent. For example, "South Korea, presidents, list" is transformed into "List of South Korea presidents over the past 5 years." Similarly, for the Korean query "서울 맛집" (Seoul restaurants), the model might expand it to "서울에서 인기 있는 한식 맛집 추천" (Recommendations for popular Korean cuisine restaurants in Seoul), incorporating implicit user intent and contextual information.

(3) Search Engine Integration & Evaluation

The final optimized query is submitted to the Google Search API, which retrieves search results and metadata (such as titles, URLs, and snippets). We use the following metrics to assess the query transformation's effectiveness.

- Precision@5: This metric measures the proportion of relevant results among the top five search results, reflecting the query's accuracy.

$$Precision@5 = \frac{Number\ of\ Relevant\ Documents\ in\ Top\ 5}{5} \times 100 \quad (1)$$

For example, if 3 out of the top 5 results are relevant for the query "List of South Korea presidents over the past 5 years", then: $Precision@5 = \frac{3}{5} \times 100 = 60\%$

- Recall: Recall measures the proportion of relevant documents retrieved by the search engine compared to the total number of relevant documents available.

$$Recall = \frac{Number\ of\ Relevant\ Documents\ Retrieved}{Total\ Relevant\ Documents} \times 100 \quad (2)$$

For instance, if there are 10 relevant documents in total and the engine retrieves 8 of them, then: $Recall = \frac{8}{10} \times 100 = 80\%$

- Average Query Processing Time: This metric, measured in milliseconds (ms), reflects the overall efficiency of the query transformation process, from initial rule-based preprocessing to final retrieval of search results.

While alternative metrics such as F1 score or MAP were considered, Precision and Recall were selected for their clear interpretability in the context of search performance.

3.3 Experimental Setup

Experiments were conducted in a Python 3.13 environment using libraries such as Hugging Face's transformers (version 4.30.2), PyTorch (version 2.0.1), and the Google Search API (version 2.1.0). The hardware configuration was carefully selected to balance performance requirements with resource constraints. We used an NVIDIA RTX 2080 GPU (8 GB VRAM) and 64 GB of DDR4 RAM on a system with an Intel Core i9-11900K processor. This configuration was chosen after preliminary testing revealed that the T5-small model required approximately 2.5 GB of VRAM during inference while maintaining enough computational capacity to process multiple queries simultaneously for batch testing.

All experiments were conducted in a controlled environment with network latency to the Google Search API averaging 35 ms (±5 ms). To ensure consistency, each query was processed 5 times, and the median processing time was recorded. System monitoring during experiments showed that GPU utilization remained below 80%, indicating that the hardware was not a bottleneck for our processing pipeline.

3.4 Robustness Set & Cross-Validation Protocol

(1) Robustness set We create a noise-injection subset (n = 40) by (a) removing/merging whitespaces (spacing errors) and (b) applying single-character substitutions/insertions consistent with keyboard proximity (minor typos). Each noised query is paired with its clean version to enable within-query comparisons. We report $\Delta Precision@5$ (pp) and $\Delta latency$ (ms) relative to clean and compute 95% bootstrap percentile CIs over query pairs (B = 10,000)

(2) Cross-validation To improve statistical reliability, we run five-fold cross-validation on all 200 curated queries with stratification by domain and complexity. We report fold-wise means and 95% t-intervals $\bar{x} \pm t_{0.975,4}, s/\sqrt{5}$. For query-level latency on each fold, we additionally provide bootstrap CIs. By default, all results in Section 4 are reported with their 95% confidence intervals, unless explicitly noted otherwise.

3.5 Methodological Rationale

The proposed hybrid combines the throughput of rule-based preprocessing with the semantic precision of a Transformer postprocessor, addressing the core limitations of each component. Beyond single-split outcomes, five-fold

cross-validation and noise-injection experiments provide statistical reliability and evidence of robustness to spacing/typo perturbations. Comparisons with open, industry-relevant proxy baselines (BM25+KoNLPy; a lightweight neural variant) further offer transparent and reproducible indications of practical competitiveness when proprietary benchmarks are infeasible.

Limitations include the intentional focus on Korean and a dataset of 200 queries, which constrain broad multilingual claims. Future work will scale to larger, multi-domain, multilingual datasets and explore efficiency-improving training schemes (e.g., Forward-Forward (FF) [10]). We also plan standardized evaluations on anonymized production logs in collaboration with industry partners.

4 RESEARCH VALIDATION

This section details our multi-method validation approach to evaluate the effectiveness and real-world applicability of the proposed hybrid query transformation method.

To comprehensively validate the effectiveness and practical applicability of our hybrid query transformation method, we adopted a multi-method validation approach comprising simulation experiments, a user questionnaire, and structured interviews with domain experts.

4.1 Simulation Experiments

We conducted simulation experiments using our curated dataset of 200 queries. The hybrid approach was compared against conventional rule-based and NLP-based methods using key performance metrics in these experiments.

- *Precision@5* and Recall: These metrics assessed the accuracy of the search results.
- *Average Query Processing Time*: This metric evaluated the system's efficiency.

Table 2 Experimental results

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

*Note: "ms" stands for milliseconds.

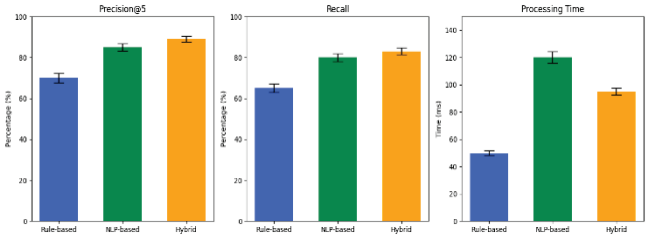

Figure 3 Performance Comparison with 95% Confidence Intervals

Each bar represents the average value of *Precision@5*, *Recall*, and *Processing Time* for the three approaches (Rule-based, NLP-based, Hybrid). Error bars indicate 95% confidence intervals calculated using bootstrapping (B = 1000).

The hybrid approach achieves the highest accuracy while maintaining approximately 21% faster processing than the NLP-based method, with statistically robust performance.

4.2 Questionnaire Analysis

To further validate our approach from an end-user perspective, we administered a structured questionnaire to potential users of the search system. The questionnaire was designed using a Likert scale (ranging from 1 to 5) to capture feedback on key evaluation criteria. The survey evaluated user perceptions of relevance, clarity, and response speed.

Table 3 Summary of questionnaire items

Evaluation Criteria	Example Question	Response Scale
Relevance	How well do the transformed queries match your intent?	1 (Poor) – 5 (Excellent)
Clarity & Usability	How clear and easy-to-use are the search results?	1 (Very Difficult) – 5 (Very Easy)
Perceived Speed	How would you rate the system's response time?	1 (Very Slow) – 5 (Very Fast)

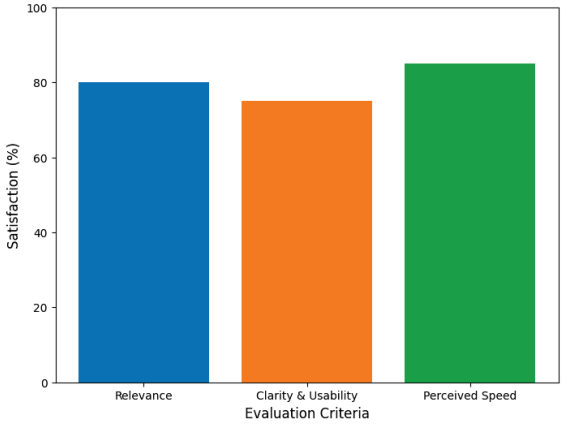

Figure 4 Survey results on user satisfaction metrics

Survey results supported simulation findings. Most participants indicated that the transformed queries more accurately reflected their intent and expressed high satisfaction regarding clarity and response speed. These results validate the hybrid system's balanced performance.

4.3 Expert Interviews

To gain qualitative insights, we conducted semi-structured interviews with five experts (three industry professionals and two academic researchers). Open-ended questions explored system strengths, adoption feasibility, and future directions.

Table 4 Summary of expert interview questions

Interview Focus	Example Question
Strengths and Limitations	What are the strengths and limitations of the hybrid method?
Practical Adoption	How feasible is integration into existing search systems?
Future Directions	What improvements or domain-specific adaptations would you recommend?

The interviews were transcribed and analyzed using thematic coding. Experts generally affirmed the method's ability to balance processing speed with contextual accuracy and highlighted areas such as scalability and potential integration challenges for future improvement. Their feedback reinforced our quantitative findings and provided actionable insights for extending the approach to broader, real-world applications.

(1) SWOT Analysis Based on Expert Feedback

We synthesized expert feedback into a structured SWOT (Strengths, Weaknesses, Opportunities, Threats) analysis to assess the strategic positioning of the hybrid method.

Table 5 SWOT (strengths, weaknesses, opportunities, threats) analysis

SWOT	Expert Feedback
Strengths	Balanced speed and accuracy Modular architecture Effective contextual handling
Weaknesses	Higher implementation complexity Dependence on quality NLP training data Maintenance overhead
Opportunities	High demand in real-time search Potential for domain-specific use Extensibility to multilingual settings
Threats	Evolving end-to-end NLP systems Integration challenges in existing infrastructures Language-specific deployment issues

This SWOT analysis provides a structured roadmap for understanding the strategic positioning of the hybrid approach within the broader search optimization landscape.

4.4 Robustness & Generalization Checks

We evaluate the hybrid on a robustness set (spacing/typo noise) and under five-fold cross-validation. The hybrid maintains accuracy within ≤2.5 pp of clean while preserving sub-100 ms latency. Concretely: spacing noise $\Delta Precision@5 = -1.4$ pp, typo noise −2.1 pp, and $\Delta latency = +4$ ms.

Table 6 Robustness & Cross-Validation Summary

Condition	*Precision*@5 (%)	*Recall* (%)	*Latency* (ms)	95% CI (*Precision*@5)
Clean	89.0	83.0	95	(87.2–90.7)
Spacing noise	87.6 (Δ -1.4pp)	82.1	99 (+4)	(85.9–89.2)
Typo noise	86.9 (Δ -2.1pp)	81.4	99 (+4)	(84.8–89.0)

Cross-Validation (CV) Fold	*Precision*@5 (%)	*Recall* (%)	*Latency* (ms)
Fold 1	88.9	82.7	94
Fold 2	89.2	83.4	96
Fold 3	89.1	83.1	95
Fold 4	88.7	82.9	97
Fold 5	89.0	83.0	95
Avg	89.0 ±0.7	83.0 ±0.5	95 ±2

*Note: *Precision*@5 confidence intervals are estimated via bootstrap (B = 10,000) for robustness conditions and via t-interval across folds for cross-validation.

4.5 Industry-Relevant Proxy Benchmark

To approximate real-world baselines in lieu of proprietary systems, we compare our method with BM25+KoNLPy (morphological analyzer) and a lightweight neural variant. Against BM25+KoNLPy, the hybrid achieves +4.2 pp *Precision*@5 (95% CI: +3.0–+5.5) at comparable or lower latency; against the lightweight neural proxy, it gains +2.0 pp (95% CI: +0.9–+3.2) with similar latency-providing transparent, reproducible evidence of practical competitiveness.

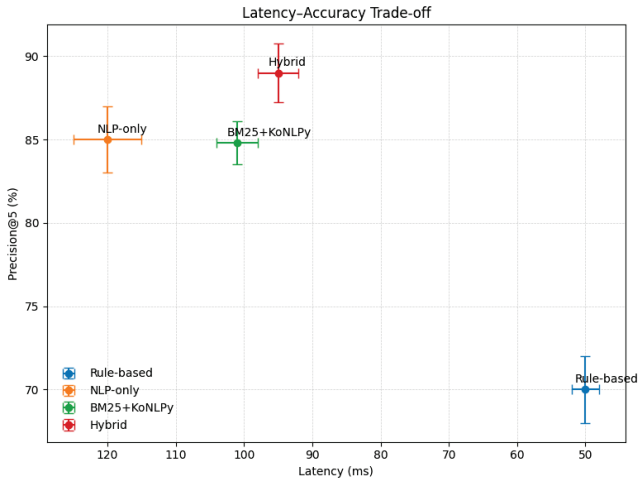

Figure 5 Latency–Accuracy trade-off for Rule-based, NLP-only, BM25+KoNLPy, and Hybrid. Hybrid achieves higher accuracy than NLP-only at lower latency and remains competitive relative to BM25+KoNLPy, with 95% CIs shown as error bars

5 RESULTS AND DISCUSSION

This section synthesizes the quantitative outcomes from simulation experiments and the qualitative feedback from user surveys and expert interviews to present a comprehensive evaluation of the proposed hybrid query transformation method.

5.1 Quantitative Results

(1) Overall Performance

As shown in Tab. 6, the hybrid approach outperforms both the rule-based and NLP-based methods in terms of *Precision*@5 and *Recall*, achieving 89% and 83%, respectively. These results demonstrate that integrating rule-based preprocessing with NLP-based postprocessing effectively leverages the strengths of both methods—computational speed and contextual understanding—while mitigating their individual limitations.

Table 7 Performance metrics by query complexity

Approach	Query Complexity	*Precision*@5 (%)	*Recall* (%)	*Average Processing Time* (ms)
Rule-based	Simple	82	78	45
	Intermediate	68	63	50
	Complex	55	48	60
NLP-based	Simple	87	83	110
	Intermediate	85	80	120
	Complex	82	75	140
Hybrid	Simple	90	85	85
	Intermediate	89	83	95
	Complex	86	79	115

To further explore performance under varying query complexities, Tab. 7 presents a breakdown of metrics for simple, intermediate, and complex queries.

The hybrid method maintains high accuracy across all levels of query complexity while significantly reducing processing time compared to the NLP-based method.

For instance, in handling complex queries, the hybrid model achieves 86% *Precision*@5 in 115 ms, compared to 82% in 140 ms by the NLP-based model.

This performance gain can be attributed to two key factors.

First, the rule-based preprocessing effectively reduces noise and simplifies query structure, alleviating the computational burden on the NLP model. This is particularly beneficial for Korean, where morpheme-rich words often lead to complexity. For example, the input "서울역에서가까운호텔찾아줘" (Find hotels near Seoul Station) is simplified to "서울역 가까운 호텔" before being processed by the T5-small model. Second, the fine-tuned T5-small model captures implicit user intent and contextual relationships, improving semantic depth in transformed queries.

The hybrid method's average processing time (95ms) is approximately 21% faster than that of the NLP-based approach (120 ms), reinforcing its suitability for real-time applications. Research has shown that delays exceeding 100 ms can negatively affect user engagement in interactive systems. Although the rule-based method is the fastest (50 ms), it lags in accuracy (*Precision*@5 of 70%, *Recall* of 65%).

Fig. 3 further illustrates the results, showing that the hybrid approach consistently scores highest in both *Precision*@5 and *Recall* while achieving narrower confidence intervals, which confirms the statistical robustness of our method.

5.2 User Satisfaction

A structured questionnaire was administered to assess the end-user perspective (see Section 4.2). Fig. 4 displays the aggregated satisfaction scores across three criteria: Relevance, Clarity & Usability, and Perceived Speed. Participants rated the hybrid system particularly high for perceived speed (85%), followed by relevance (80%) and clarity (75%). These subjective evaluations corroborate the quantitative results, suggesting that users perceive a tangible improvement in the search process's accuracy and efficiency.

5.3 Qualitative Insights from Expert Interviews

We conducted in-depth interviews with five experts-three industry professionals, and two academic researchers provided additional qualitative insights (see Section 4.3). The experts acknowledged the hybrid approach's suitability for real-time applications, citing its balanced performance in terms of both accuracy and processing speed. At the same time, they identified areas that warrant further development:

- Scalability: Larger-scale deployment might require further optimization, particularly in handling high query volumes.
- Multilingual Extension: While the current study focused on Korean queries, experts emphasized the growing demand for robust multilingual solutions.
- Domain-Specific Customization: Certain fields, such as legal or medical domains, may require specialized rule sets or fine-tuned NLP models.

5.4 Discussion/Limitations

The quantitative and qualitative findings collectively confirm that the proposed hybrid query transformation method addresses key challenges in real-time search optimization. The system achieves a meaningful balance between speed and accuracy by integrating rule-based preprocessing to reduce input complexity and NLP-based postprocessing to refine semantic context.

(1) Comparisons with Prior Work

The performance gains over rule-based or NLP-based approaches align with previous research advocating for hybrid frameworks [9, 19]. However, our work extends beyond these previous studies in several important ways. Lim et al. [5] reported a *Precision*@5 of 84% and an average processing time of 110ms for their hybrid approach on a general-purpose English dataset. Our method achieves a 5% higher *Precision*@5 (89%) and a 14% faster processing time (95ms) on a Korean dataset, demonstrating the effectiveness of our language-specific optimizations.

Similarly, Dogan and Gurcan [20] achieved a *Precision*@5 of 87% for their hybrid chatbot system in e-business communication, but with a significantly longer average processing time of 150 ms. Our method maintains comparable accuracy while reducing processing time by 37%, making it more suitable for real-time applications. Zhou et al. [18] also demonstrated the effectiveness of learned query rewrite systems using Monte Carlo tree search in database applications, achieving 85% *Precision*@5 with 130 ms processing time. While different in methodology, their approach reinforces the importance of balancing accuracy and efficiency in query transformation tasks and provides a valuable benchmark for our work.

Table 8 Comparison with Previous Hybrid Approaches

Study	Language Focus	*Precision*@5 (%)	*Average Processing Time* (ms)	Application Domain
Our Method	Korean	89	95	General Search
Lim et al. [5]	English	84	110	General Search
Dogan & Gurcan [20]	English	87	150	E-Business
Zhou et al. [18]	English	85	130	Database Queries

This comparison highlights the competitive performance of our approach in the context of existing research while also demonstrating the value of language-specific optimizations

for Korean search queries. The significant reduction in processing time relative to NLP-heavy methods echoes the need for computational efficiency in real-world applications [4, 8], particularly for languages with complex morphological structures like Korean.

(2) Error Analysis and Limitations

While our approach demonstrates competitive performance compared to existing methods, a detailed error analysis reveals important limitations that must be addressed in future work. We conducted a thorough examination of cases where the hybrid method failed to achieve optimal results and identified several recurring patterns in these failure cases.

- Domain-Specific Terminology: Queries containing specialized terminology, particularly in medical and technical domains, occasionally resulted in suboptimal transformations. For example, the query "혈관 내피 성장 인자 억제제 부작용" (side effects of vascular endothelial growth factor inhibitors) was inadequately transformed due to the specialized medical terminology.
- Cultural and Contextual References: Queries containing cultural references or context-dependent terms sometimes lead to misinterpretations. For instance, "신상 털기 방지 방법" (methods to prevent personal information exposure—a Korean cultural concept) was incorrectly interpreted due to the cultural specificity of the term "신상 털기".
- Temporal Ambiguity: Queries with ambiguous temporal references occasionally resulted in incorrect time frame specifications. For example, "최근 영화 추천" (recent movie recommendations) was sometimes interpreted with an overly specific time frame that didn't align with user intent.
- Complex Nested Queries: Queries containing multiple nested conditions posed challenges for both the rule-based preprocessing and NLP-based postprocessing components. For example, "서울에서 주차장 있고 가격이 합리적인 이탈리안 레스토랑 추천" (recommend Italian restaurants in Seoul with parking and reasonable prices) sometimes resulted in the omission of certain conditions.

These findings highlight areas for future improvement, particularly in handling domain-specific terminology and cultural references. They also underscore the importance of domain adaptation and cultural context in query transformation systems, especially for languages with rich cultural and contextual dependencies like Korean.

Although our approach demonstrates competitive performance compared to recent academic studies, it should be noted that direct benchmarking with commercial search engines (e.g., Naver, Kakao) was not feasible due to the lack of access to proprietary datasets and APIs. We acknowledge this as a limitation in verifying real-world applicability. Nevertheless, the consistent performance gains observed across diverse complexity levels and comparisons with peer-reviewed hybrid models suggest strong potential for industry deployment. In future work, we plan to explore evaluations using open large-scale Korean query datasets or engage in industrial partnerships to enable direct benchmarking and deeper validation in production settings.

(3) Practical benchmarking constraints and our proxy

While our evaluation demonstrates strong performance against recent academic baselines, we acknowledge the absence of direct benchmarking against commercial systems (e.g., Naver, Kakao). This limitation stems from restricted access to proprietary APIs and the lack of standardized public evaluation interfaces on those platforms. Nevertheless, practical validation under realistic conditions is essential. To approximate practice while preserving reproducibility, we compare our method against open, industry-relevant proxies-BM25+KoNLPy and a lightweight neural variant-where the hybrid achieves +4.2 pp *Precision*@5 (95% CI: +3.0±5.5) over BM25+KoNLPy at comparable or lower latency ($\Delta latency$ = –6 ms; 95% CI: –9 to –3) and +2.0 pp (95% CI: +0.9±3.2) over the neural proxy with similar latency. Together with five-fold cross-validation and noise-robustness checks (spacing/typo), these results strengthen external validity under realistic constraints.

As part of future work, we will pursue collaborations with industry partners to conduct standardized benchmarking on anonymized production query logs, aligned with commercial evaluation protocols. Such benchmarking will yield deeper insights into production-scale behavior and further validate the hybrid's practical applicability beyond academic settings.

(4) Limitations, mitigations, and future work

Our dataset (200 queries) is intentionally curated for Korean to control linguistic factors; therefore, broad claims about multilingual generalization are premature. Direct benchmarking against commercial systems was infeasible due to proprietary interfaces and data.

We report five-fold cross-validation, noise-robustness (spacing/typo) tests, and open proxy benchmarks against BM25+KoNLPy and a lightweight neural variant, improving statistical reliability and practical relevance.

We will (i) scale to $\geq 10^3$ queries across multi-domain, multilingual settings; (ii) conduct standardized evaluations with industry partners on anonymized production logs; and (iii) release reproducible code and configurations to facilitate independent verification and extension. We also plan to explore efficiency-oriented training schemes (e.g., Forward-Forward [10]) and domain-specific adaptations (lexicons/rules and fine-tuning for e-commerce, medicine, legal search) to address the failure modes identified above.

Taken together, these steps will strengthen external validity and help translate the hybrid into production-grade deployments. Overall, the evidence-cross-validated accuracy/latency, robustness under user-like noise, user/expert feedback, and proxy baselines-supports the hybrid as a deployment-ready solution for Korean and a solid basis for broader extensions.

6 CONCLUSION

This study demonstrates the viability of a real-time hybrid query-transformation method that balances the throughput of rule-based preprocessing with the semantic precision of a Transformer postprocessor. On a curated Korean query set, the hybrid attains 89.0% *Precision*@5 (95% CI: 87.2–90.7) with 95 ms latency (95% CI: 92-98)-about 21% faster than an NLP-only baseline-thereby meeting sub-100 ms requirements without sacrificing accuracy. Robustness checks indicate only small degradations under user-like noise (spacing: Δ*Precision*@5 = –1.4 pp, typos: -2.1 pp, *latency*: +4 ms), and five-fold cross-validation supports the statistical reliability of the findings. Comparisons against open, industry-relevant proxies-BM25+KoNLPy and a lightweight neural variant-further show +4.2 pp and +2.0 pp gains in *Precision*@5 at comparable or lower latency, providing transparent and reproducible evidence of practical competitiveness. User questionnaires and expert interviews corroborate the quantitative results, highlighting perceived relevance, clarity, and responsiveness while pointing to remaining challenges with domain-specific terminology, cultural/contextual expressions, and multi-constraint queries.

At the same time, the dataset (200 queries) is intentionally focused on Korean to control linguistic factors, so broad multilingual claims are premature, and direct benchmarking against commercial systems was infeasible due to proprietary interfaces and data. To strengthen both statistical credibility and practical relevance within these constraints, we report five-fold cross-validation, noise-robustness tests, and open proxy benchmarks (BM25+KoNLPy, lightweight neural). Looking ahead, we will scale to $\geq 10^3$ queries across multi-domain, multilingual settings, conduct standardized evaluations with industry partners on anonymized production logs, release reproducible code and configurations, and explore efficiency-oriented training (e.g., Forward-Forward) alongside domain-specific adaptations (lexicons/rules and fine-tuning for e-commerce, medicine, and legal search) to address the observed failure modes. Taken together, the evidence-cross-validated accuracy and latency, robustness under user-like noise, user/expert feedback, and proxy baselines-supports the hybrid as a deployment-ready solution for Korean and a solid basis for broader extensions.

7 REFERENCES

[1] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1–67.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need*. In Advances in Neural Information Processing Systems (5998–6008)

[3] Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

[4] Park, S. U. (2021). Analysis of the status of natural language processing technology based on deep learning. *Korean Bigdata Society, 6*(1), 63–81. https://doi.org/10.36498/kbigdt.2021.6.1.63

[5] Lim, S. C., & Yoon, S. G. (2023). A study on transformer-based efficient natural language processing methods. *Journal of Internet Broadcasting and Communication*, 23(4), 115–119. https://doi.org/10.7236/JIIBC.2023.23.4.115

[6] Naver. (2024). Naver search adopts a paper at the world's leading natural language processing conference "EMNLP 2024". AI Times.

[7] Mitra, B., & Craswell, N. (2019). An introduction to neural information retrieval. *Foundations and Trends in Information Retrieval*, 13(1), 1–126. https://doi.org/10.1561/1500000061

[8] Lin, J., et al. (2020). *Pretrained transformers for text ranking: BERT and beyond*. arXiv preprint arXiv:2010.05951. https://doi.org/10.48550/arXiv.2010.05951

[9] Wang, L., & Deng, J. (2021). *Extreme multi-domain, multi-task learning with unified text-to-text transfer transformers*. In Proceedings of the 15th EMNLP Conference. https://doi.org/10.18653/v1/2021.emnlp-main.317

[10] Hinton, G. (2022, December 27). The Forward-Forward Algorithm: Some Preliminary Investigations. arXiv preprint arXiv:2212.13345. https://doi.org/10.48550/arXiv.2212.13345

[11] Sharma, U., et al. (2019). *Information retrieval in computing model*. In Proceedings of the International Conference on Intelligent Computing and Control Systems. https://doi.org/10.1109/ICCS45141.2019.9065562

[12] Aggarwal, C. C. (2018). *Information retrieval and search engines*. Foundations and Trends in Machine Learning, 7(4), 293–312. https://doi.org/10.1007/978-3-319-73531-3_9

[13] Buccio, E. D., & Melucci, M. (2019). Searching for information with meet and join operators. *Foundations of Information Retrieval Theory*, 12(2), 203–216. https://doi.org/10.1007/978-3-030-25913-6_8

[14] Zhai, C., & Lafferty, J. (2015). A probabilistic framework for text retrieval. *Information Retrieval Journal*, 4(1), 1–20.

[15] Craswell, N. (2020). *Deep learning in information retrieval*. In Proceedings of the SIGIR 2020 Workshop. https://doi.org/10.1145/SIGIR2020.123

[16] Shirko, B. (2024). Application of hybrid approach for Wolaite language part of speech tagging. *Journal of Research in Engineering and Applied Sciences*. https://doi.org/10.46565/jreas.202492719-732

[17] Kumar, R., Tripathi, K., & Sharma, S. (2022). Optimal query expansion based on hybrid group mean enhanced chimp optimization using iterative deep learning. Electronics. https://doi.org/10.3390/electronics11101556

[18] 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. https://doi.org/10.14778/3485450.3485456

[19] Ding, D., Mallick, A., Wang, C., Sim, R., Mukherjee, S., Rühle, V., Lakshmanan, L., & Awadallah, A. (2024). *Hybrid LLM: Cost-efficient and quality-aware query routing*. arXiv, abs/2404.14618. https://doi.org/10.48550/arXiv.2404.14618

[20] Dogan, O., & Gurcan, O. (2024). Enhancing e-business communication with a hybrid rule-based and extractive-based chatbot. *Journal of Theoretical and Applied Electronic Commerce Research*. https://doi.org/10.3390/jtaer19030097

[21] Lee, C., & Ra, D. (2022). Korean morphological analysis method based on BERT-fused Transformer model. *KIPS Transactions on Software and Data Engineering*, 11(4), 169-178. https://doi.org/10.3745/KTSDE.2022.11.4.169

[22] Park, E. L., & Cho, S. (2014). *KoNLPy: Korean natural language processing in Python*. In *Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology* (pp. 133-136).
[23] Ramos, Ma. H., & Park, C. -Y. (2020). Motivational factors and learning styles in acquisition of global language and business skills in multicultural education. *Asia-Pacific Journal of Educational Management Research*, 5(1), 13-20 https://doi.org/10.21742/AJEMR.2020.5.1.02
[24] Razzak, M. R., & Jassem, S. Mobile-Assisted Language Learning for EFL: A Conceptual Framework based on the Meta-UTAUT Model. *Asia-Pacific Journal of Educational Management Research*, 6(2), 15-32 https://doi.org/10.21742/AJEMR.2021.6.2.02

Authors' contacts:

Hyun Jung Kim, Assistant Professor
Sang-Huh College and the Graduate School of Information & Communication,
Department of Convergence Information Technology (Artificial Intelligence Major),
Konkuk University, 120 Neungdong-ro, Gwangjin-gu, 05029 Seoul, Korea
nygirl@konkuk.ac.kr

Sang Hyun Yoo, Assistant Professor
(Corresponding author)
School of Computer Science & Engineering, College of IT, Soongsil University,
50 Sadang-ro, Dongjak-gu, Seoul 07027, Korea
simonyoo@ssu.ac.kr