SHRAG: A Framework for Combining Human-Inspired Search with RAG

RDGENAI 2025 1st Place Winner: A New RAG Paradigm Discovered Through Failure

Hyunseok Ryu[1] Wonjune Shin[1] Hyun Park[2]

Gwangju Institute of Science and Technology (GIST)
omnyx2@gmail.com

November 14, 2025

[1] Equal contribution

Outline

Framework

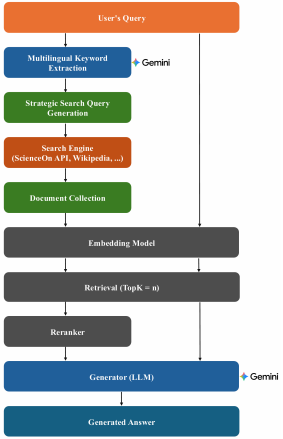

Figure: The name of framework: Human like but not 100%.

Introduction: What Problem Were We Solving?

Here is the Challenge Goal.

RDGENAI Challenge

Develop a RAG system that retrieves scientific papers to answer complex questions and generates responses that are accurate, fluent, and properly cited.

The assignment we received was rather vague. Also, several stringent constraints complicated the problem:

- **Multilingual Setting:** Queries provided in either English or Korean.

Introduction: What Problem Were We Solving?

Here is the Challenge Goal.

RDGENAI Challenge

Develop a RAG system that retrieves scientific papers to answer complex questions and generates responses that are accurate, fluent, and properly cited.

The assignment we received was rather vague. Also, several stringent constraints complicated the problem:

- **Multilingual Setting:** Queries provided in either English or Korean.
- **Strict Formatting:** Responses must follow a Title-Introduction-Body-Conclusion structure.

Introduction: What Problem Were We Solving?

Here is the Challenge Goal.

RDGENAI Challenge

Develop a RAG system that retrieves scientific papers to answer complex questions and generates responses that are accurate, fluent, and properly cited.

The assignment we received was rather vague. Also, several stringent constraints complicated the problem:

- **Multilingual Setting:** Queries provided in either English or Korean.
- **Strict Formatting:** Responses must follow a Title-Introduction-Body-Conclusion structure.
- **Composite Evaluation:** Retrieval accuracy, response quality, and runtime all factored into scoring.

Development: Initial Setback

As RAG practitioners, we began with textbook approaches.

Development: Initial Setback

As RAG practitioners, we began with textbook approaches.

- Naive RAG pipeline
- Hybrid retrieval (Sparse + Dense)

Development: Initial Setback

As RAG practitioners, we began with textbook approaches.

- Naive RAG pipeline
- Hybrid retrieval (Sparse + Dense)

The outcome?

Development: Initial Setback

As RAG practitioners, we began with textbook approaches.

- Naive RAG pipeline
- Hybrid retrieval (Sparse + Dense)

The outcome?

Disappointing Failure

Scores(0.275) fell far short of expectations.

Development: Initial Setback

As RAG practitioners, we began with textbook approaches.

- Naive RAG pipeline
- Hybrid retrieval (Sparse + Dense)

The outcome?

Disappointing Failure

Scores(0.275) fell far short of expectations.

This marked the start of our genuine research: **What went wrong?**

Development: Four Hypotheses on Failure Causes

We manually inspected all 50 evaluation questions and formulated four hypotheses:

1. **Hypothesis 1 (Formatting):** “Could it be that Gemini is getting points deducted for failing to adhere to the format?”

Development: Four Hypotheses on Failure Causes

We manually inspected all 50 evaluation questions and formulated four hypotheses:

1. **Hypothesis 1 (Formatting):** “Could it be that Gemini is getting points deducted for failing to adhere to the format?”
2. **Hypothesis 2 (Evaluation):** “Do METEOR/BLEU metrics have issues with dealing with Korean responses?”

Development: Four Hypotheses on Failure Causes

We manually inspected all 50 evaluation questions and formulated four hypotheses:

1. **Hypothesis 1 (Formatting):** “Could it be that Gemini is getting points deducted for failing to adhere to the format?”
2. **Hypothesis 2 (Evaluation):** “Do METEOR/BLEU metrics have issues with dealing with Korean responses?”
3. **Hypothesis 3 (Data):** “Does the task require a real-time search in addition to the information in test.csv?”

Development: Four Hypotheses on Failure Causes

We manually inspected all 50 evaluation questions and formulated four hypotheses:

1. **Hypothesis 1 (Formatting):** “Could it be that Gemini is getting points deducted for failing to adhere to the format?”
2. **Hypothesis 2 (Evaluation):** “Do METEOR/BLEU metrics have issues with dealing with Korean responses?”
3. **Hypothesis 3 (Data):** “Does the task require a real-time search in addition to the information in test.csv?”
4. **Hypothesis 4 (Complexity):** “Are the questions actually complex multi-hop reasoning tasks?”

Development: Second Attempt and Greater Failure

We implemented solutions targeting each hypothesis:

- (H1) $\rightarrow$ Few-shot prompting (0.267 $\rightarrow$ 0.259)
- (H2) $\rightarrow$ Multilingual embeddings (mGTE) (0.250 $\rightarrow$ 0.267)
- (H3) $\rightarrow$ Real-time ScienceON search integration (0.563 $\rightarrow$ 0.344)
- (H4) $\rightarrow$ Query decomposition logic (0.267 $\rightarrow$ 0.267)

Development: Second Attempt and Greater Failure

We implemented solutions targeting each hypothesis:

- (H1) $\rightarrow$ Few-shot prompting (0.267 $\rightarrow$ 0.259)
- (H2) $\rightarrow$ Multilingual embeddings (mGTE) (0.250 $\rightarrow$ 0.267)
- (H3) $\rightarrow$ Real-time ScienceON search integration (0.563 $\rightarrow$ 0.344)
- (H4) $\rightarrow$ Query decomposition logic (0.267 $\rightarrow$ 0.267)

The result?

Development: Second Attempt and Greater Failure

We implemented solutions targeting each hypothesis:

- (H1) $\rightarrow$ Few-shot prompting (0.267 $\rightarrow$ 0.259)
- (H2) $\rightarrow$ Multilingual embeddings (mGTE) (0.250 $\rightarrow$ 0.267)
- (H3) $\rightarrow$ Real-time ScienceON search integration (0.563 $\rightarrow$ 0.344)
- (H4) $\rightarrow$ Query decomposition logic (0.267 $\rightarrow$ 0.267)

The result?

Critical Failure

Scores deteriorated further; we were lost in a maze.

Development: Second Attempt and Greater Failure

We implemented solutions targeting each hypothesis:

- (H1) $\rightarrow$ Few-shot prompting (0.267 $\rightarrow$ 0.259)
- (H2) $\rightarrow$ Multilingual embeddings (mGTE) (0.250 $\rightarrow$ 0.267)
- (H3) $\rightarrow$ Real-time ScienceON search integration (0.563 $\rightarrow$ 0.344)
- (H4) $\rightarrow$ Query decomposition logic (0.267 $\rightarrow$ 0.267)

The result?

Critical Failure

Scores deteriorated further; we were lost in a maze.

(With only 50 samples, quantitative validation was infeasible.)

Comparison: AI and Big Data in Neurological Medical Imaging

SAI's: Artificial Intelligence and Big Data in Neurology

Introduction
Recent technological developments have enabled the collection and analysis of vast amounts of clinical and imaging data in neurology. These resources—including electronic health records, population registries, and advanced neuroimaging databases—offer unprecedented opportunities to enhance diagnosis and treatment of neurological disorders.

Main Body
Key advances include the deployment of machine learning models that automate lesion detection and segmentation on MRI and CT scans, improving both speed and consistency compared to manual interpretation. Deep learning frameworks have demonstrated high accuracy in classifying complex patterns associated with stroke, multiple sclerosis, Alzheimer's disease, and other conditions, leveraging convolutional neural networks trained on large annotated datasets. Integration of real-world data (RWD) and multi-modal analytics has supported the development of predictive models for disease progression and treatment response, while large-scale collaborative initiatives are standardizing data formats and promoting federated learning to protect patient privacy. Despite these successes, challenges persist: variability in image acquisition protocols compromises model generalizability, data heterogeneity and missing labels hamper robust training, and ensuring data quality and provenance remains difficult. Ethical concerns around informed consent, algorithmic bias, and data security further complicate deployment in clinical settings, and regulatory frameworks are still evolving to address

Ours: Advances and Challenges... (Structured Format)

The application of Artificial Intelligence (AI) and Big Data in medical imaging, particularly in neurology, has seen significant advancements, offering transformative potential for diagnosis, prognosis, and treatment. These technological strides have opened up access to vast datasets and enabled sophisticated analytical capabilities. However, alongside these opportunities, the field faces considerable challenges that need to be addressed for successful and ethical integration into clinical practice.
Main Advances: Recent technological progress has facilitated access to a multitude of datasets across various dimensions of neurology, providing a rich foundation for AI and Big Data applications. Deep learning-powered AI has demonstrated remarkable progress in image recognition tasks, proving its feasibility for diverse applications in medical imaging. Specifically in neuroimaging, AI can significantly improve clinical practice by:

- **Enhanced Diagnosis and Prognosis:** Aiding in the detection of brain metastases and predicting the treatment response of brain tumors. AI also augments clinicians' intelligence in making more informed diagnostic, prognostic, and treatment decisions.
- **Advanced Image Analysis:** Generating parametric maps from dynamic contrast-enhanced MRI and enhancing radiomics research through the extraction of salient features from input images.
- **Improved Image Quality:** Enhancing image quality via AI-based image reconstruction and reducing motion artifacts, leading to clearer and more reliable diagnostic images.
- **Real-World Data Utilization** [illegible] Real-World Data (RWD) and

Pivot: Ground Truth via Human Evaluation

In the end, we chose the most primitive but most accurate method,‘Human Judging’. Three of our team members re-opened all the problems and the hundreds of results that were generated.

Pivot: Ground Truth via Human Evaluation

In the end, we chose the most primitive but most accurate method,'Human Judging'. Three of our team members re-opened all the problems and the hundreds of results that were generated.
This revealed that most of our hypotheses were incorrect:

- **Truth 1 (Formatting):** Few-shot prompting performed worse than zero-shot (due to markdown noise).

Pivot: Ground Truth via Human Evaluation

In the end, we chose the most primitive but most accurate method,'Human Judging'. Three of our team members re-opened all the problems and the hundreds of results that were generated.
This revealed that most of our hypotheses were incorrect:

- **Truth 1 (Formatting):** Few-shot prompting performed worse than zero-shot (due to markdown noise).
- **Truth 2 (Multilingual):** mGTE embeddings worked well for retrieving correct documents.

Pivot: Ground Truth via Human Evaluation

In the end, we chose the most primitive but most accurate method,'Human Judging'. Three of our team members re-opened all the problems and the hundreds of results that were generated.
This revealed that most of our hypotheses were incorrect:

- **Truth 1 (Formatting):** Few-shot prompting performed worse than zero-shot (due to markdown noise).
- **Truth 2 (Multilingual):** mGTE embeddings worked well for retrieving correct documents.
- **Truth 3 (Search):** Real-time search quality was awful (engine limitations).

Pivot: Ground Truth via Human Evaluation

In the end, we chose the most primitive but most accurate method,'Human Judging'. Three of our team members re-opened all the problems and the hundreds of results that were generated.
This revealed that most of our hypotheses were incorrect:

- **Truth 1 (Formatting):** Few-shot prompting performed worse than zero-shot (due to markdown noise).
- **Truth 2 (Multilingual):** mGTE embeddings worked well for retrieving correct documents.
- **Truth 3 (Search):** Real-time search quality was awful (engine limitations).
- **Truth 4 (Complexity):** Over 85% of questions were single-hop.

Pivot: Revised Conclusion

The puzzle pieces finally aligned:

- We were solving for multi-hop $\rightarrow$ but the task was single-hop.
- We enforced few-shot generation control $\rightarrow$ zero-shot was superior.
- We confirmed that multilingual embeddings functioned effectively.

Pivot: Revised Conclusion

The puzzle pieces finally aligned:

- We were solving for multi-hop $\rightarrow$ but the task was single-hop.
- We enforced few-shot generation control $\rightarrow$ zero-shot was superior.
- We confirmed that multilingual embeddings functioned effectively.

Key Insight

The bottleneck was not LLM generation, it was **how to search exact documents with multilingual user query**.

Conclusion: SHRAG: A Framework for Combining Human-Inspired Search with RAG

At this juncture, we radically shifted direction.

Conclusion: SHRAG: A Framework for Combining Human-Inspired Search with RAG

At this juncture, we radically shifted direction.

Core Question

“In a multilingual setting with a low-quality search engine, how can we reliably retrieve trustworthy documents?”

Conclusion: SHRAG: A Framework for Combining Human-Inspired Search with RAG

At this juncture, we radically shifted direction.

Core Question

"In a multilingual setting with a low-quality search engine, how can we reliably retrieve trustworthy documents?"

Inspiration came from human search behavior:

Conclusion: SHRAG: A Framework for Combining Human-Inspired Search with RAG

At this juncture, we radically shifted direction.

Core Question

"In a multilingual setting with a low-quality search engine, how can we reliably retrieve trustworthy documents?"

Inspiration came from human search behavior:

- When there is a question human's are generally search by keyword which is unknown first and search with combination of keywords.
- Also when peoples are also checking the keyword was right.
- Searching "Charles Darwin" *or* "찰스 다윈" *or* "Father of evolution"

Overall Framework

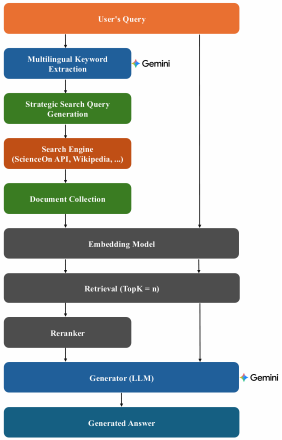

Figure: An illustration explaining the system's overall structure and components.

Conclusion: SHRAG 5-Step Pipeline

1. **Multilingual Keyword Extraction**
   - Prompt LLM to extract English and Korean keywords separately.

Conclusion: SHRAG 5-Step Pipeline

1. **Multilingual Keyword Extraction**
   - Prompt LLM to extract English and Korean keywords separately.
2. **Strategic Query Generation**
   - Combine all keywords with OR operators:
     `SQ = (Charles OR Darwin OR 찰스 OR 다윈 OR ...)`

Conclusion: SHRAG 5-Step Pipeline

1. **Multilingual Keyword Extraction**
   - Prompt LLM to extract English and Korean keywords separately.
2. **Strategic Query Generation**
   - Combine all keywords with OR operators:
     `SQ = (Charles OR Darwin OR 찰스 OR 다윈 OR ...)`
3. **Document Retrieval**
   - Broad retrieval using OR query (high recall).

Conclusion: SHRAG 5-Step Pipeline

1. **Multilingual Keyword Extraction**
   - Prompt LLM to extract English and Korean keywords separately.
2. **Strategic Query Generation**
   - Combine all keywords with OR operators:
     `SQ = (Charles OR Darwin OR 찰스 OR 다윈 OR ...)`
3. **Document Retrieval**
   - Broad retrieval using OR query (high recall).
4. **Multilingual Re-ranking**
   - Compute cosine similarity between query and documents using mGTE; select top-5 (high precision).

Conclusion: SHRAG 5-Step Pipeline

1. **Multilingual Keyword Extraction**
   - Prompt LLM to extract English and Korean keywords separately.
2. **Strategic Query Generation**
   - Combine all keywords with OR operators:
     `SQ = (Charles OR Darwin OR 찰스 OR 다윈 OR ...)`
3. **Document Retrieval**
   - Broad retrieval using OR query (high recall).
4. **Multilingual Re-ranking**
   - Compute cosine similarity between query and documents using mGTE; select top-5 (high precision).
5. **Structured Answer Generation**
   - Feed top-5 documents to Gemini in zero-shot mode.

Award and Additional Validation

- Our proposed framework (SHRAG) achieved **1st place** in the **ScienceON AI Challenge (SAI Challenge)**.
- To further verify the robustness and generalization performance of our model, we conducted **additional validation**.
- We utilized the **Miracle dataset** for this subsequent validation.
- To properly evaluate the retrieval performance on this dataset, we defined and applied a **new metric** suitable for its characteristics: the **Query Success Rate (QSR)**.

Query Success Rate (QSR)

Definitions

- Q: The set of all evaluation queries ($|Q|$: total number of queries).
- $A_i = \{a_{1i}, a_{2i}, \ldots, a_{ki}\}$: The set of k_i known relevant documents for query q_i.
- S_i: The set of documents retrieved by the search for query q_i.

Successful Query

- A query q_i is considered successful if its retrieved set S_i contains at least one relevant document.
- Formally: $\mathbf{A_i} \cap \mathbf{S_i} \neq \emptyset$

QSR Formal Definition

$$\text{QSR} = \frac{|\{q_i \in Q \mid A_i \cap S_i \neq \emptyset\}|}{|Q|} * 100 \tag{1}$$

Result (SHRAG)

When real-time search on the Wikipedia. There could be another title with similar and valid content for question.

Core Strategy 1: Why OR, Not AND?

Principle: “Search broadly (OR), then refine precisely (mGTE)”

Core Strategy 1: Why OR, Not AND?

Principle: “Search broadly (OR), then refine precisely (mGTE)”

If AND were used?

Missing even one keyword excludes the correct document entirely.

With OR?

All relevant candidates are retrieved, giving mGTE the chance to select the correct one.

Core Strategy 1: Why OR, Not AND?

Principle: “Search broadly (OR), then refine precisely (mGTE)”

If AND were used?

Missing even one keyword excludes the correct document entirely.

With OR?

All relevant candidates are retrieved, giving mGTE the chance to select the correct one.

Table: Query Success Rate with MIRACL Dataset

Query Language	QSR	Comments
English	100	All queries contained at least one relevant document.
Korean	88	Some queries did not yield relevant documents.
English + Korean	94	The average QSR score.

Core Strategies 2 & 3

Strategy 2: Embrace Single-Hop

Abandoning the multi-hop assumption, we focused on retrieving *the single best document*. This aligned with Truth 4 and drove score gains.

Core Strategies 2 & 3

Strategy 2: Embrace Single-Hop

Abandoning the multi-hop assumption, we focused on retrieving *the single best document*. This aligned with Truth 4 and drove score gains.

Strategy 3: Generalization (MIRACL)

“Is this ScienceON-specific?”

Core Strategies 2 & 3

Strategy 2: Embrace Single-Hop

Abandoning the multi-hop assumption, we focused on retrieving *the single best document*. This aligned with Truth 4 and drove score gains.

Strategy 3: Generalization (MIRACL)

"Is this ScienceON-specific?"

- Cross-validation on Wikipedia-based **MIRACL** dataset
- Achieved 94% QSR (probability correct document is retrieved)
- $\rightarrow$ Robust across multilingual retrieval environments

Final Remarks: A New RAG Paradigm

The strategy of

“Multilingual keyword expansion + OR-based broad search + mGTE re-ranking”

secured 1st place in the ScienceON challenge.

Final Remarks: A New RAG Paradigm

The strategy of

“Multilingual keyword expansion + OR-based broad search + mGTE re-ranking”

secured 1st place in the ScienceON challenge.
The central message of this work:

Final Remarks: A New RAG Paradigm

The strategy of

“Multilingual keyword expansion + OR-based broad search + mGTE re-ranking”

secured 1st place in the ScienceON challenge.
The central message of this work:

New RAG Paradigm

Building superior RAG systems is not solely about better embedding models.
It hinges on smarter exploitation of existing search infrastructure (e.g., Boolean search).

Thank you.