

# **RAG** Fusion

Elevating Scopus Al's generative summaries

In this high-level guide, we'll explore how Retrieval Augmented Generation (RAG) Fusion works and why it's creating such a buzz in the world of AI-powered search and content generation.



# What is RAG and why is it powerful?

Before we dive into RAG Fusion, let's understand its predecessor, retrieval augmented generation (RAG).

This clever technology, commonly used in generative AI (GenAI) tools, combines two powerful components:

# Retrieval model

Like a skilled search expert, this searches through vast datasets or knowledge bases to identify and retrieve relevant content.



# Generation model

Usually a large language model (LLM), this acts as an eloquent storyteller, crafting readable responses from the retrieved information.

LLMs are trained on enormous datasets, but RAG gives them a special superpower: the ability to ground the responses they generate in content outside their training data. In other words, they can pull content from other knowledge bases to inform their responses. When integrated correctly, this approach can significantly improve the relevance of a generated response, and ensure the LLM draws exclusively on trusted, up-to-date information.

# Limitations of traditional RAG

# Despite its advanced abilities, RAG faces several challenges:



#### **Query constraints:**

It is limited by the user's ability to formulate effective queries. This can be particularly problematic when a user is new to a topic and unfamiliar with the domain vocabulary.



#### Result ranking:

It may not always surface the most relevant documents for complex queries.



#### Single perspective:

It relies on a single search, so could potentially miss valuable, related information.



#### **Context window limitations:**

It can struggle with sifting through large amounts of retrieved information.

# Introducing RAG Fusion, the next level of AI-powered search

### This is where RAG Fusion comes in.

Developed through a collaborative effort between Scopus AI experts and the academic community, RAG Fusion is a patent-pending technology setting new benchmarks in AI-powered search. Rapidly becoming the gold standard in leading generative AI libraries like LlamaIndex and LangChain, RAG Fusion is recognized for its superior performance in exploring topics.

These user-friendly libraries offer pre-trained models, APIs, and customization tools for developers and AI enthusiasts, enabling efficient development of AI applications for tasks such as text generation and search. RAG Fusion enhances the capabilities of RAG technology, providing a seamless user experience and improved functionality.

Here's how it works:



Multi-query generation

It creates various versions of a query, including potential follow-up questions the user may not have considered.

# Step 2

Comprehensive search

It performs a vector search for each query variation, casting a wider net across the vast collection in Scopus. This creates new opportunities to find relevant information.

# Step 3

Smart ranking

The results are combined and ranked intelligently using Reciprocal Rank Fusion (RRF)\*. RRF is a method that combines search results from multiple queries, giving higher weight to items that appear near the top of many result lists. This helps ensure that the most relevant and diverse results rise to the top, balancing insights from multiple perspectives.

# Step 4

Insightful summaries

Scopus AI uses these top results to generate nuanced, comprehensive responses that maintain focus on the original user intent while incorporating broader context.

<sup>\*</sup> Reciprocal Rank Fusion (RRF) was originally proposed by Cormack, Clarke, and Buettcher in their 2009 paper 'Reciprocal Rank Fusion outperforms Condorcet and individual Rank Learning Methods' (SIGIR '09). Scopus AI builds upon this foundational work in its implementation of RAG Fusion.

# Benefits of RAG Fusion in Scopus AI

RAG Fusion enhances Scopus AI by generating comprehensive and insightful summaries from trusted scholarly content.

## Here's how it works:



#### Bridges knowledge gaps:

RAG Fusion helps users ask relevant questions, even when unfamiliar with a topic.



#### **Empowers unexpected discoveries:**

Uncovers new sources and perspectives you may have otherwise missed, opening doors to revelatory insights.



#### Enhances depth and nuance:

RAG Fusion generates responses that dive into the intricate layers of academic subjects. For instance, a query about climate change might yield insights from environmental science, economics, and public policy, providing a multifaceted understanding of the topic.



#### Offers a balanced view:

Synthesizes diverse perspectives to mitigate confirmation bias.

## Real-world impact

Since its launch in January 2024, Scopus AI's RAG Fusion-powered features have consistently maintained over 90% user satisfaction ratings. This high level of user approval demonstrates the tangible benefits RAG Fusion brings to academic research and discovery.

While RAG Fusion is a well-known technique, optimizing its integration remains a valuable skill. With Scopus AI, you have access to a top-tier team specializing in maximizing the benefits of RAG Fusion.



Contact your Elsevier representative or visit our website to learn more about integrating Scopus AI into your institution's research toolkit. elsevier.com/scopus-ai