

Retrieval-Augmented Generation



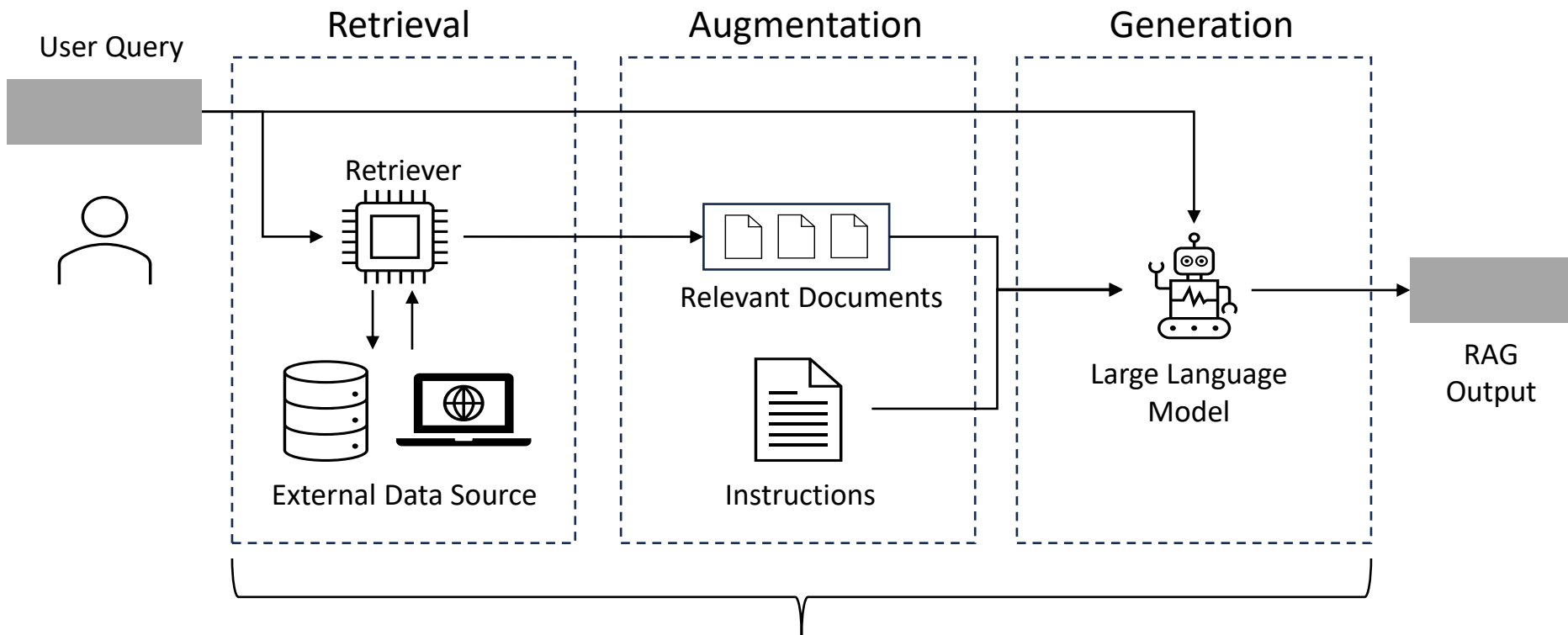
gollnickdata.de

Retrieval-Augmented Generation:

General Workflow

Retrieval-Augmented Generation

General Workflow

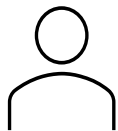


Retrieval-Augmented Generation

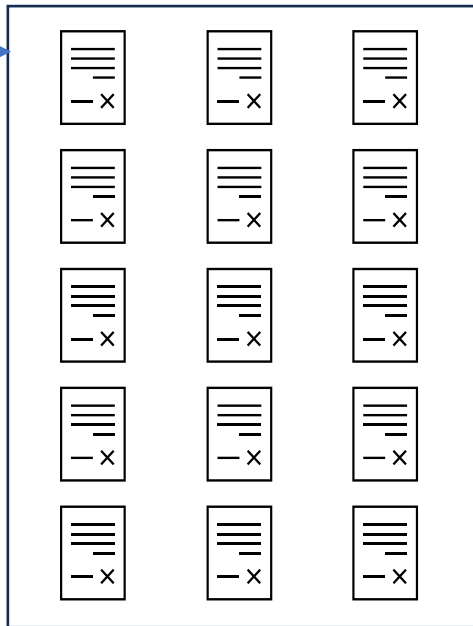
Retrieval Process

User Query

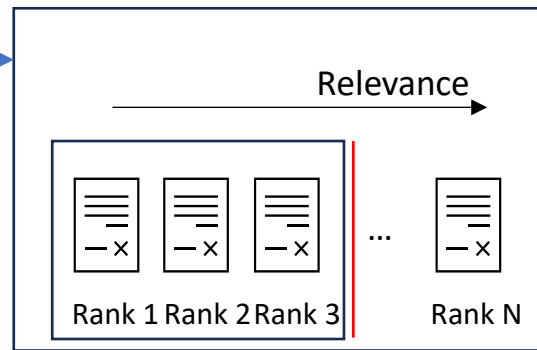
{max_results: 3}



Knowledge Source



Ranking



Returned Documents

continue with
Augmentation



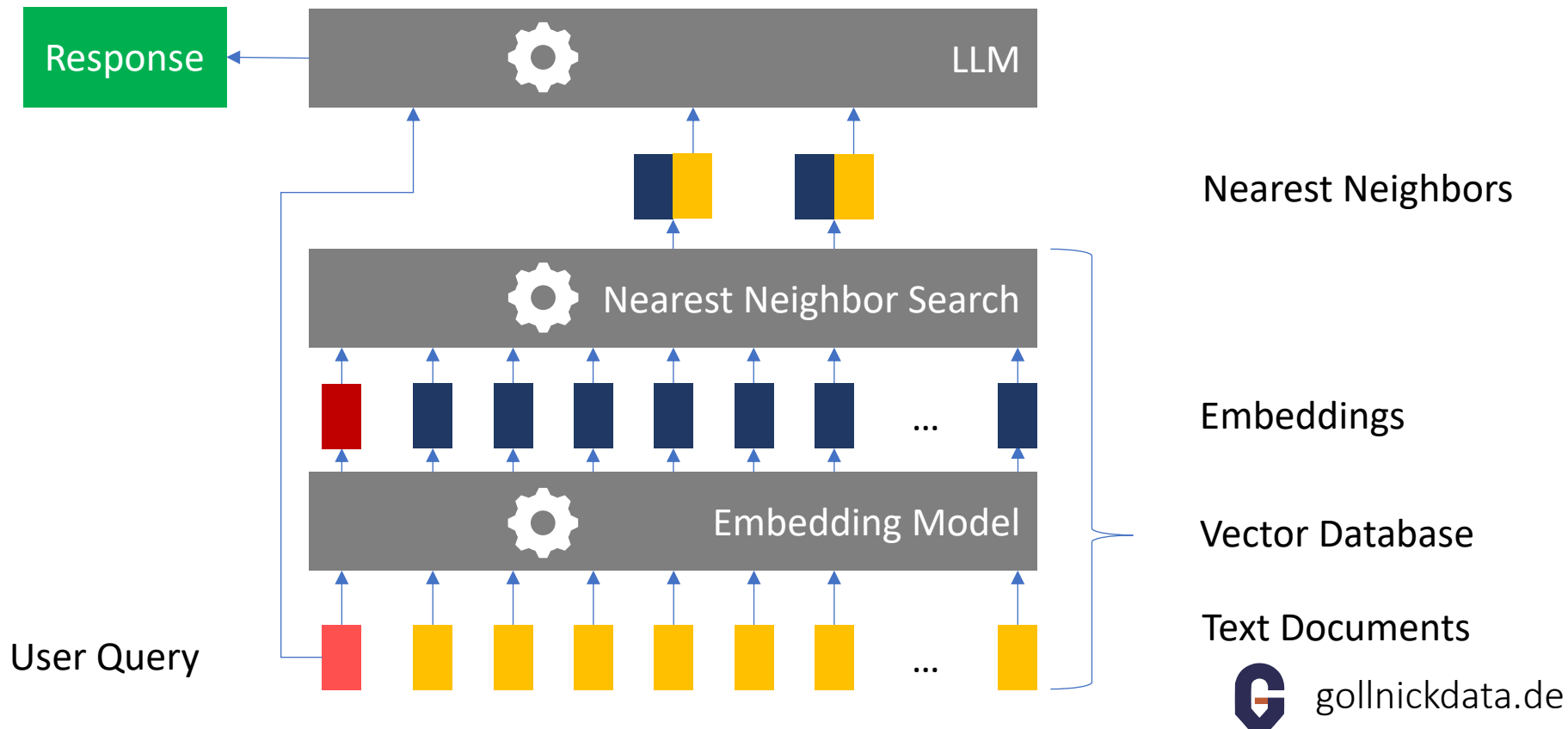
gollnickdata.de

Retrieval-Augmented Generation:

Simple RAG

Retrieval-Augmented Generation

Workflow with Vector DB Backend



Retrieval-Augmented Generation:

RAG Improvements for
Pre-Retrieval

Retrieval-Augmented Generation

RAG Improvement for Pre-Retrieval, Retrieval, and Post-Retrieval

Pre-Retrieval

- improve indexing pipeline

Retrieval

- improve retrieval process

Post-Retrieval

- improve user query

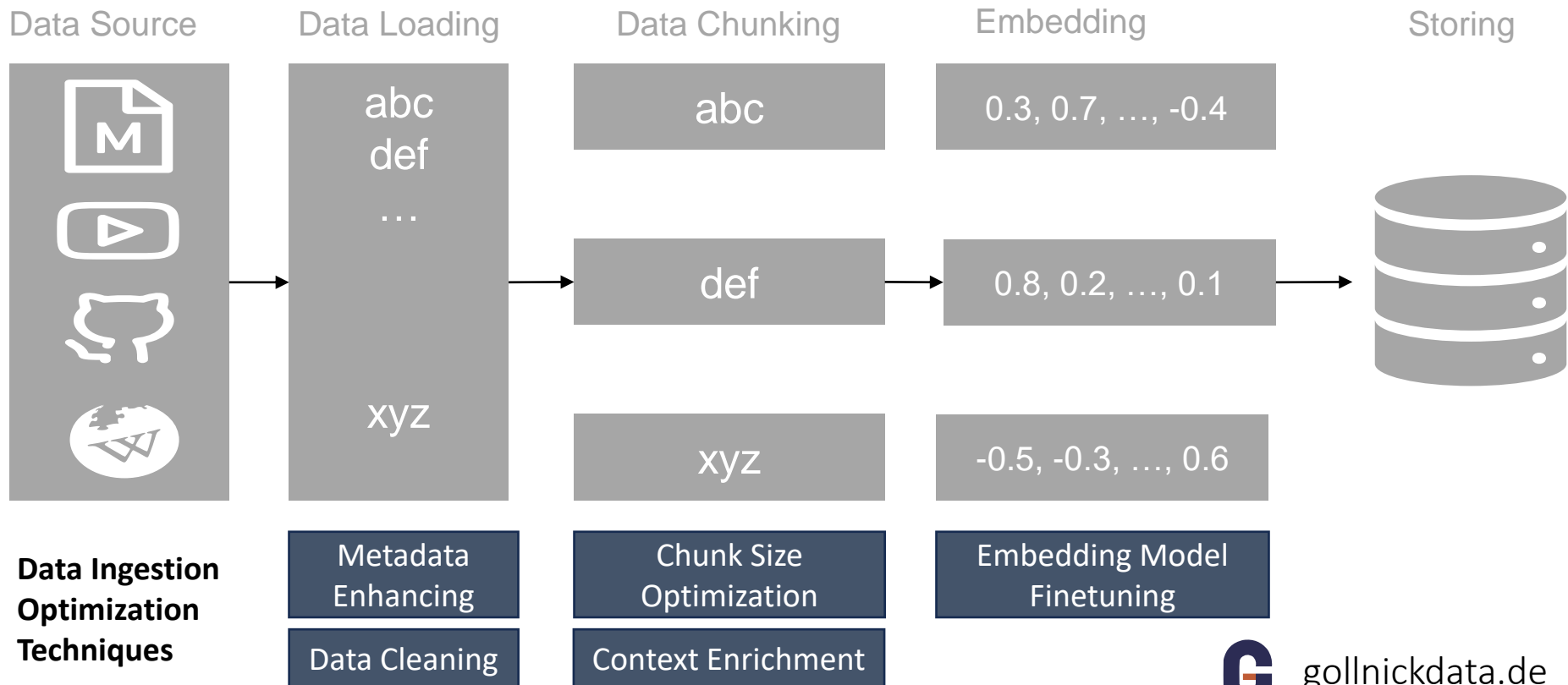


Retrieval-Augmented Generation:

RAG Improvements for
Pre-Retrieval

Retrieval-Augmented Generation

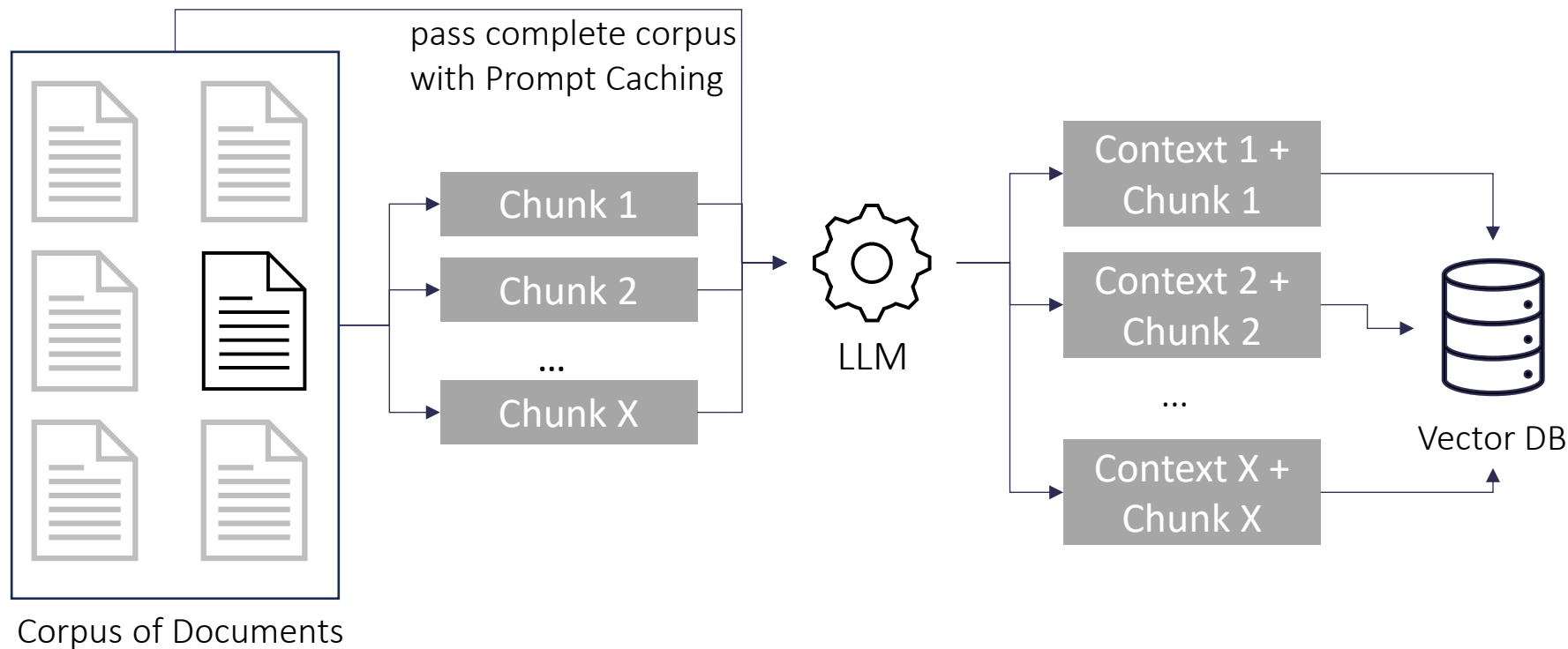
RAG Improvement Approaches Pre-Retrieval



gollnickdata.de

Retrieval-Augmented Generation

RAG Improvement for Pre-Retrieval: Context Enrichment with Contextual Retriever



own graph; adapted from <https://www.anthropic.com/news/contextual-retrieval>



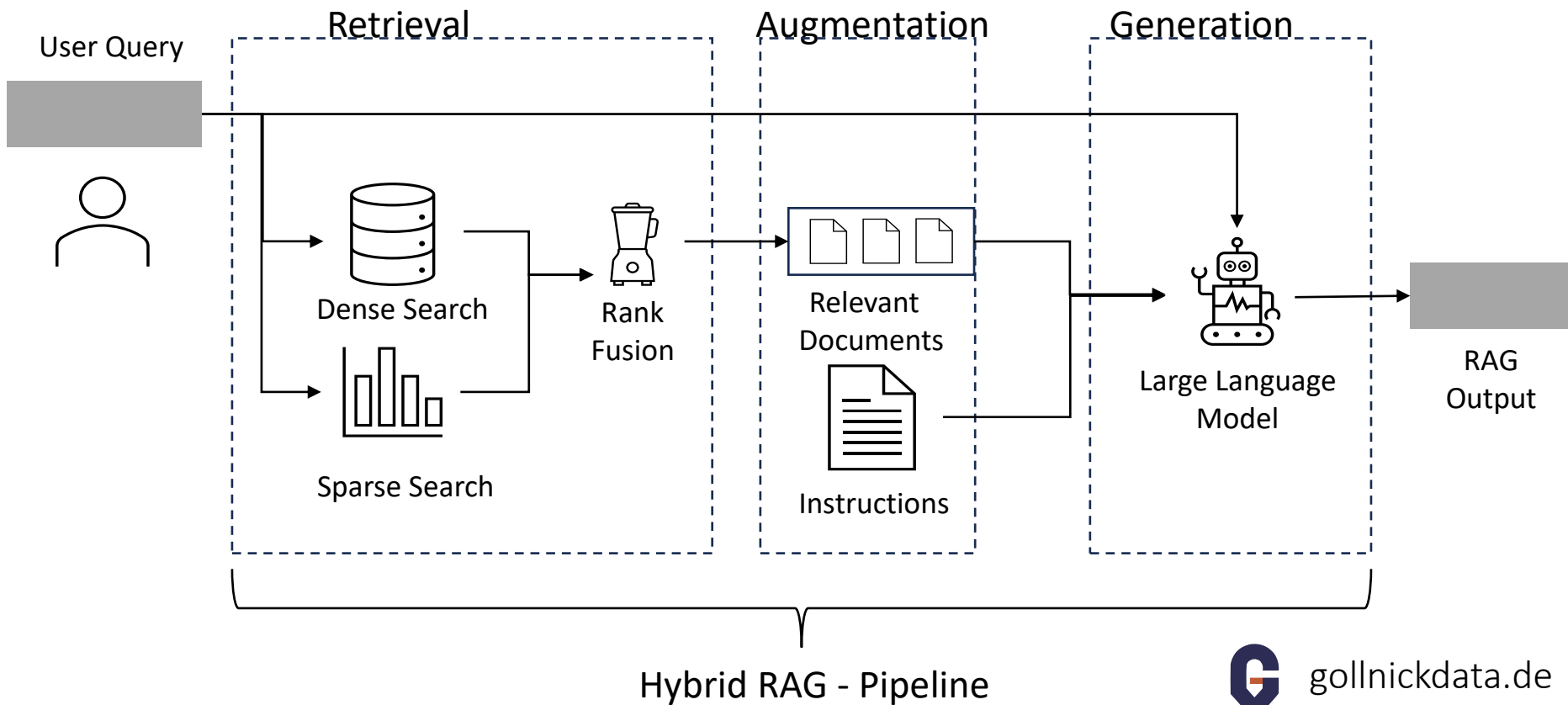
gollnickdata.de

Retrieval-Augmented Generation:

RAG Improvements for
Retrieval

Retrieval-Augmented Generation

RAG Improvement for Retrieval: Hybrid RAG



Retrieval-Augmented Generation

Reciprocal Rank Fusion

| Rank | Dense Search Result | Sparse Search Result | Reciprocal Rank |
|------|---------------------|----------------------|-----------------|
| 1 | Document D | Document A | 1/1 |
| 2 | Document A | Document C | 1/2 |
| 3 | Document C | Document B | 1/3 |
| 4 | Document B | Document D | 1/4 |

a) Reciprocal Rank Calculation

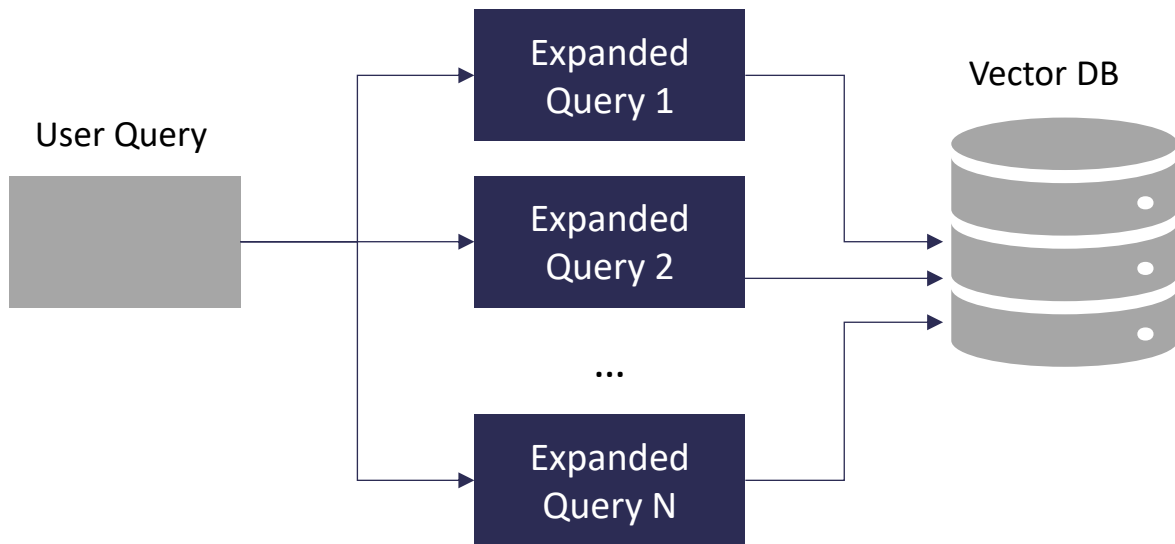
| | |
|------------|--------------------|
| Document A | $1/2 + 1/1 = 1.5$ |
| Document D | $1/1 + 1/4 = 1.25$ |
| Document C | $1/3 + 1/2 = 0.83$ |
| Document B | $1/4 + 1/3 = 0.58$ |

b) Reciprocal Rank Fusion



Retrieval-Augmented Generation

RAG Improvement for Retrieval: Query Expansion



Retrieval-Augmented Generation

RAG Improvement for Retrieval: Query Expansion

Synonym Expansion

expand with synonyms

“Climate change” →
[“global warming”,
“climate change”,
“environmental impact”]

Related Terms Expansion

related terms

“Machine learning” →
[“machine learning”,
“deep learning”,
“artificial intelligence”, “neural networks”]

Conceptual Expansion

“Renewable energy” →
[“renewable energy”,
“solar power”,
“wind energy”,
“green technology”,
“sustainable energy”]



Retrieval-Augmented Generation

RAG Improvement for Retrieval: Query Expansion

Phrase Variation Expansion

including variations of how a concept might be phrased ensures retrieval from documents that use different wording to discuss the same idea.

“Health benefits of exercise”
→
[“health benefits of physical activity”,
“exercise health benefits”,
“positive effects of exercise”]

Contextual Expansion

Including related fields or tasks

“Natural language processing” →
[“natural language processing”,
“text analysis”,
“computational linguistics”,
“language modeling”]

Temporal Expansion

Expanding for common abbreviations and alternative names →
[“COVID-19 vaccines”,
“COVID vaccines”,
“coronavirus immunization”,
“SARS-CoV-2 vaccine”]

Entity-Based Expansion

When query involves an entity, expanding with related people, products, or attributes associated with the entity can enhance retrieval quality.

“Tesla” →
[“Tesla”, “Elon Musk”,
“electrical vehicles”,
“autonomous driving”]

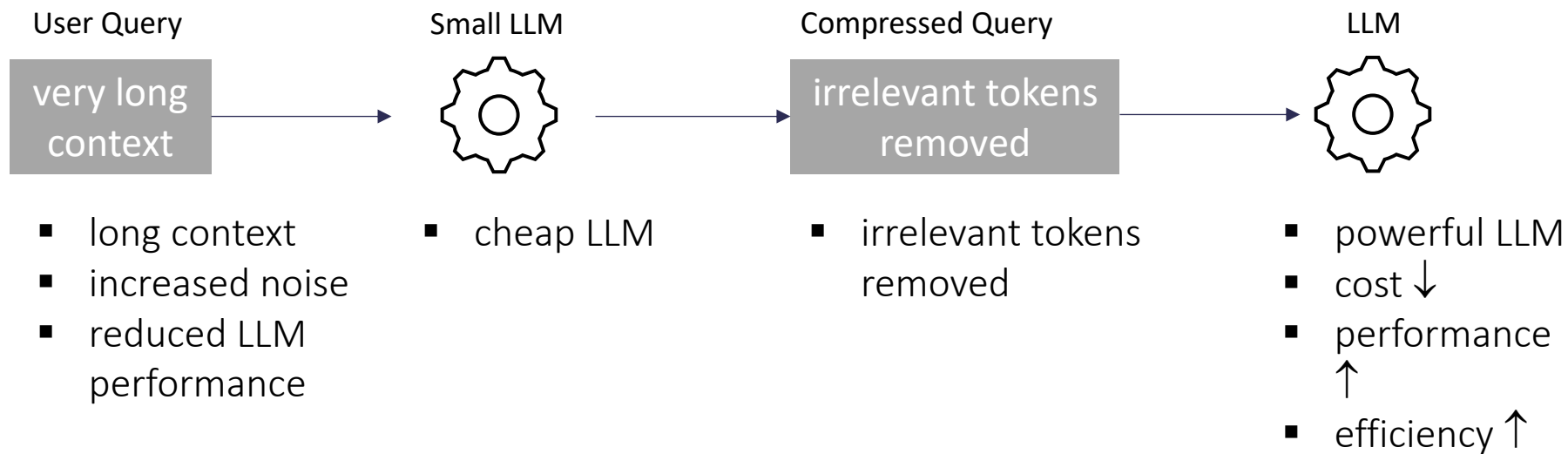


Retrieval-Augmented Generation:

RAG Improvements for
Post-Retrieval

Retrieval-Augmented Generation

RAG Improvement for Post-Retrieval: Prompt Compression



Retrieval-Augmented Generation:

RAG Alternatives

Retrieval-Augmented Generation

Prompt Caching

