

Retrieval-Augmented Generation



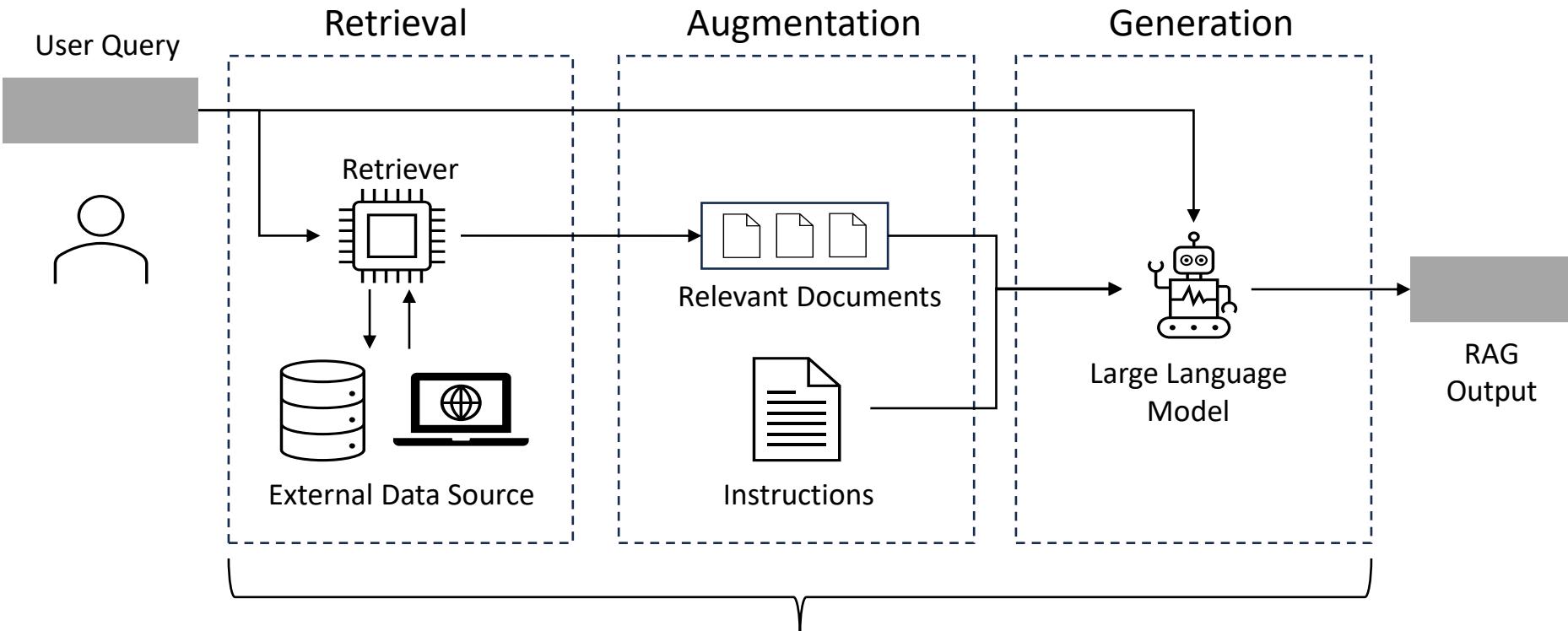
gollnickdata.de

Retrieval-Augmented Generation:

General Workflow

Retrieval-Augmented Generation

General Workflow



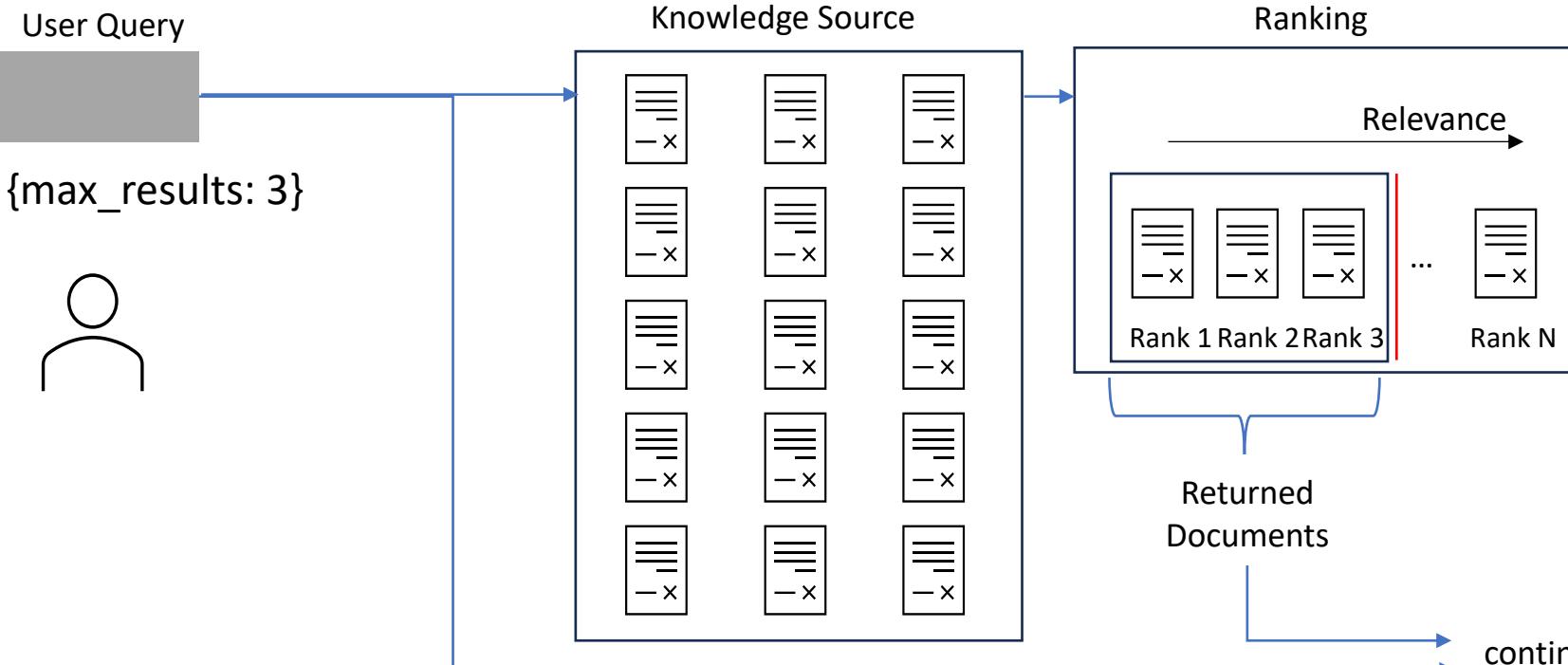
Retrieval-Augmented Generation Pipeline



gollnickdata.de

Retrieval-Augmented Generation

Retrieval Process

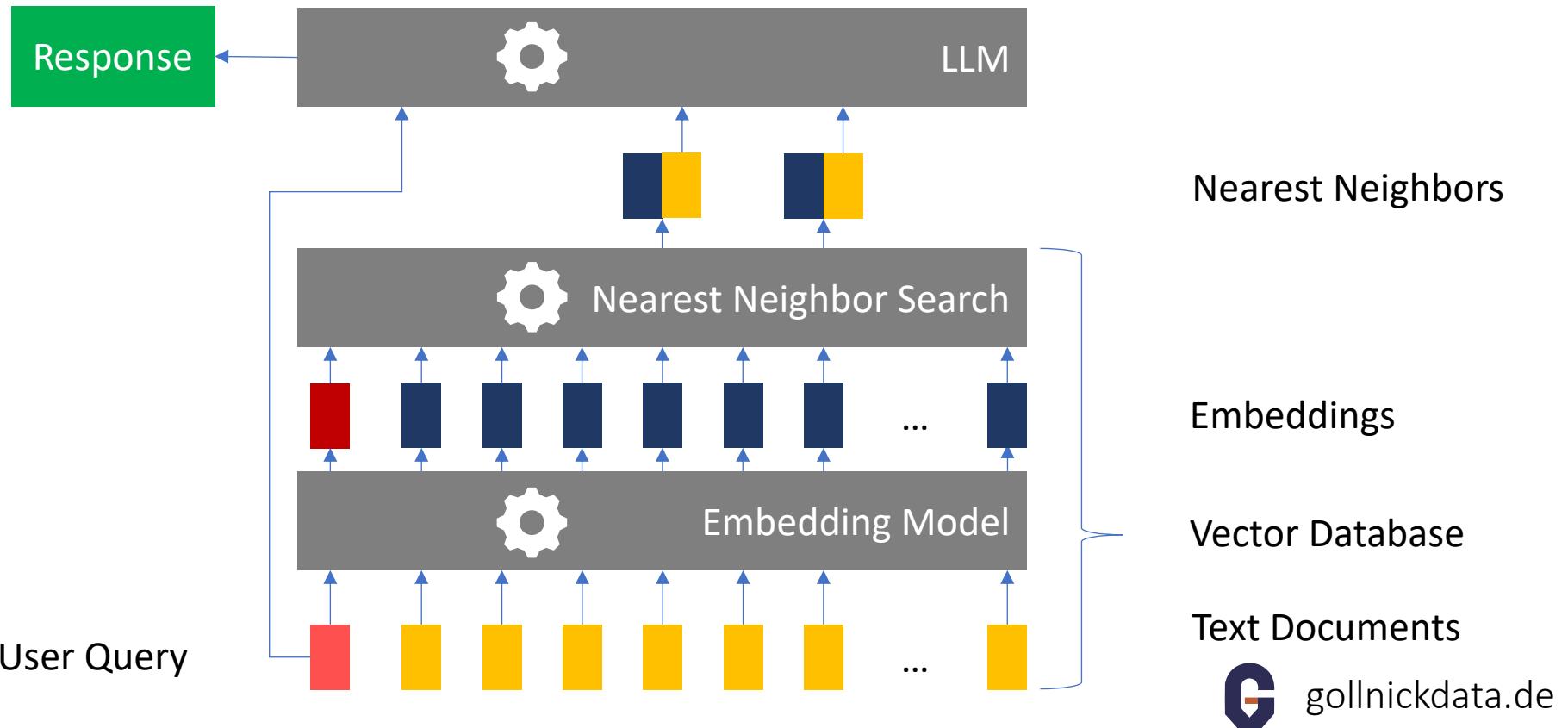


Retrieval-Augmented Generation:

Simple RAG

Retrieval-Augmented Generation

Workflow with Vector DB Backend

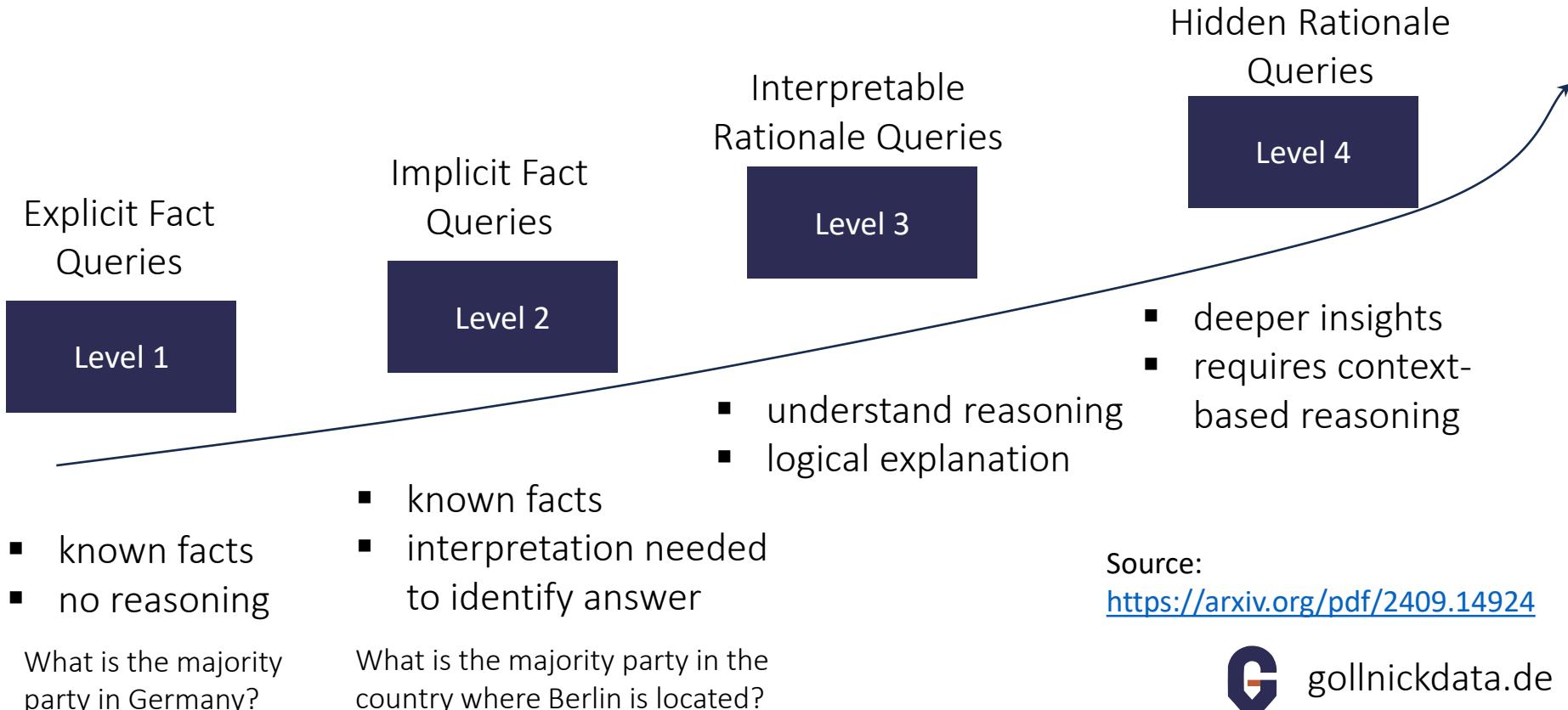


Retrieval-Augmented Generation:

RAG Improvements for
Pre-Retrieval

Retrieval-Augmented Generation

Levels of RAG



Retrieval-Augmented Generation

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

Pre-Retrieval

Retrieval

Post-Retrieval

- improve indexing pipeline
- improve retrieval process
- improve user query



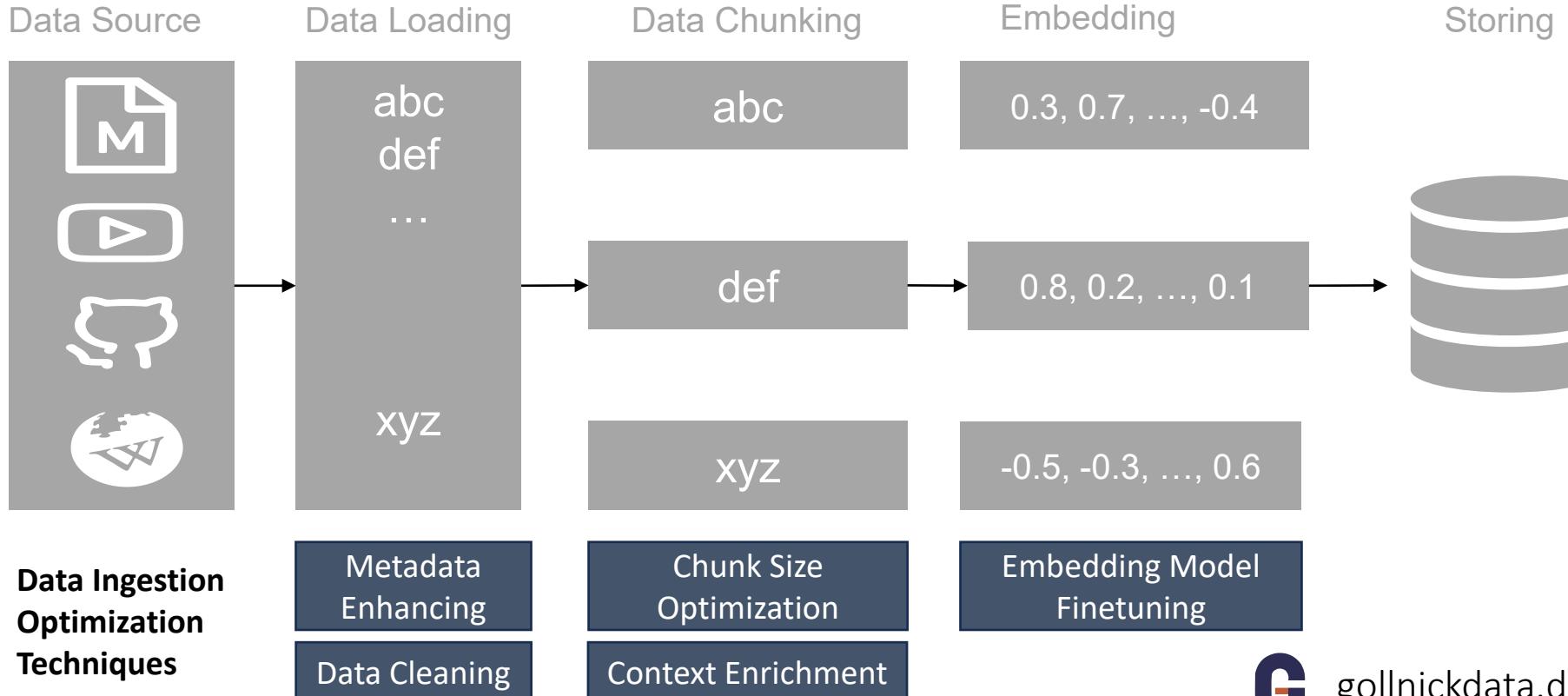
gollnickdata.de

Retrieval-Augmented Generation:

RAG Improvements for Pre-Retrieval

Retrieval-Augmented Generation

RAG Improvement Approaches Pre-Retrieval



**Data Ingestion
Optimization
Techniques**

Metadata
Enhancing

Data Cleaning

Chunk Size
Optimization

Context Enrichment

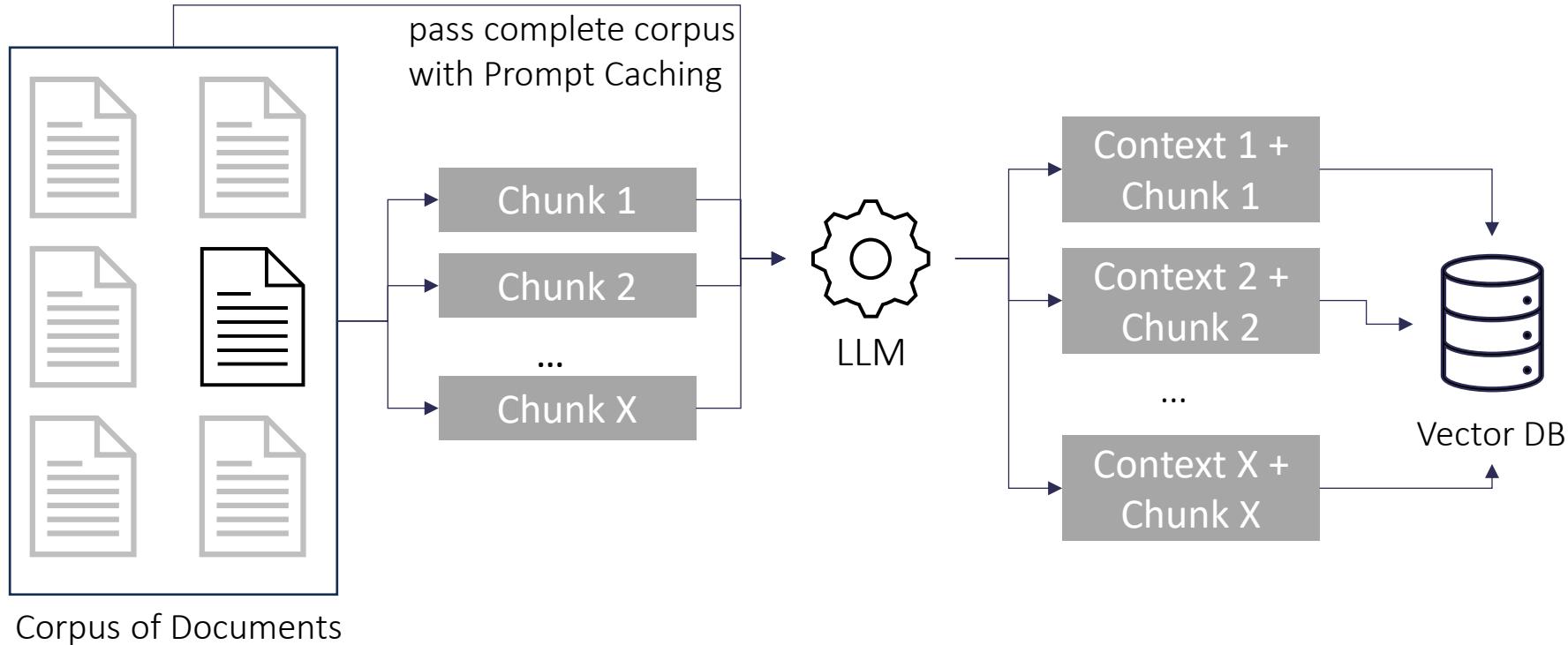
Embedding Model
Finetuning



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>



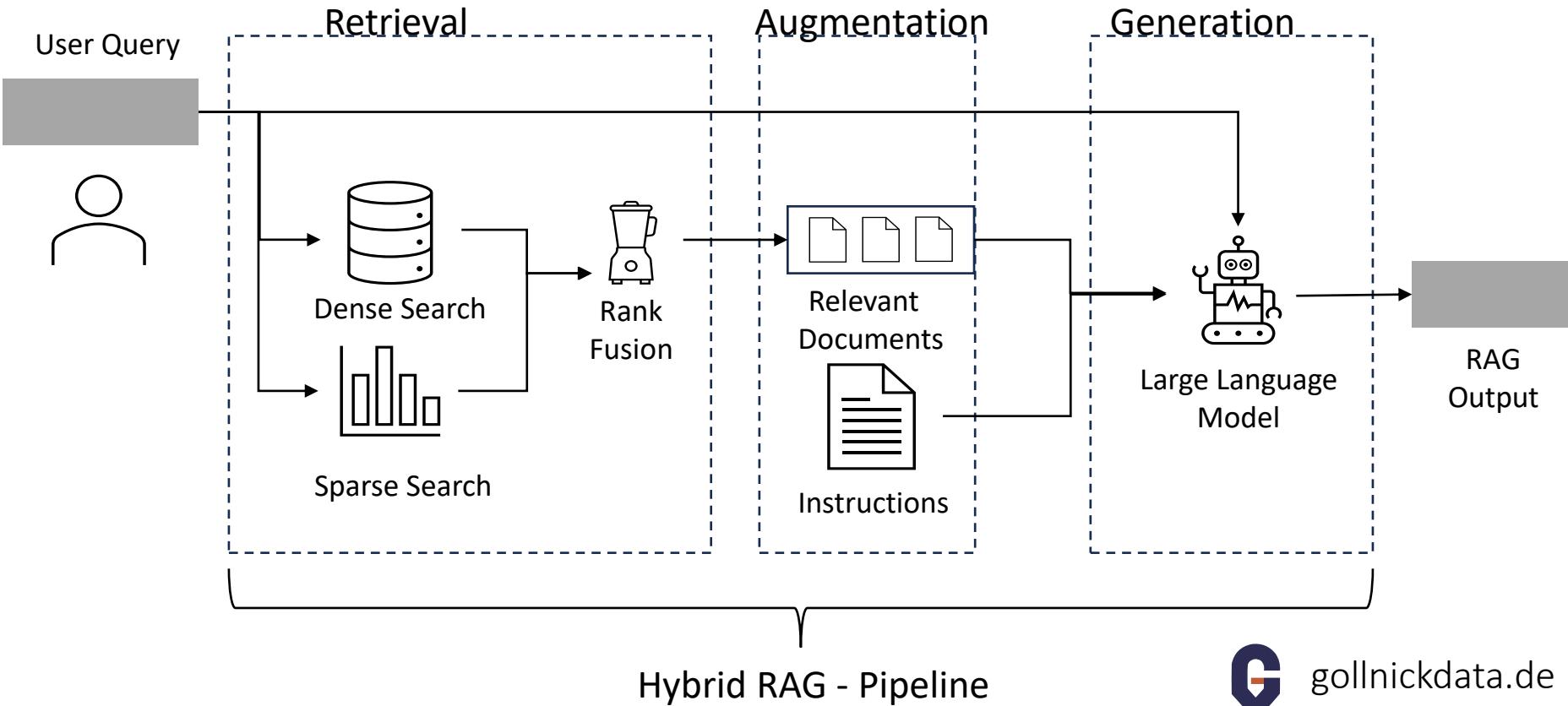
gollnickdata.de

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

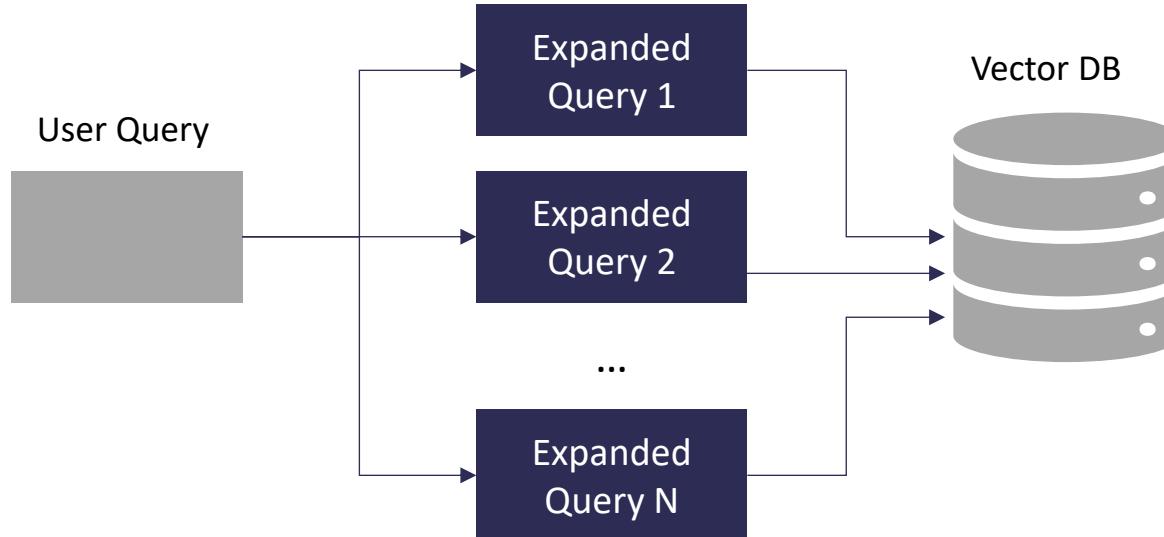
a) Reciprocal Rank Calculation

b) Reciprocal Rank Fusion



Retrieval-Augmented Generation

RAG Improvement for Retrieval: Query Expansion



gollnickdata.de

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”]



gollnickdata.de

Retrieval-Augmented Generation

RAG Improvement for Retrieval: Query Expansion

Phrase Variation Expansion	Contextual Expansion	Temporal Expansion	Entity-Based Expansion
<p>including variations of how a concept might be phrased ensures retrieval from documents that use different wording to discuss the same idea.</p> <p>“Health benefits of exercise” →</p> <p>[“health benefits of physical activity”, “exercise health benefits”, “positive effects of exercise”]</p>	<p>Including related fields or tasks</p> <p>“Natural language processing” →</p> <p>[“natural language processing”, “text analysis”, “computational linguistics”, “language modeling”]</p>	<p>Expanding for common abbreviations and alternative names →</p> <p>[“COVID-19 vaccines”, “COVID vaccines”, “coronavirus immunization”, “SARS-CoV-2 vaccine”]</p>	<p>When query involves an entity, expanding with related people, products, or attributes associated with the entity can enhance retrieval quality.</p> <p>“Tesla” →</p> <p>[“Tesla”, “Elon Musk”, “electrical vehicles”, “autonomous driving”]</p>

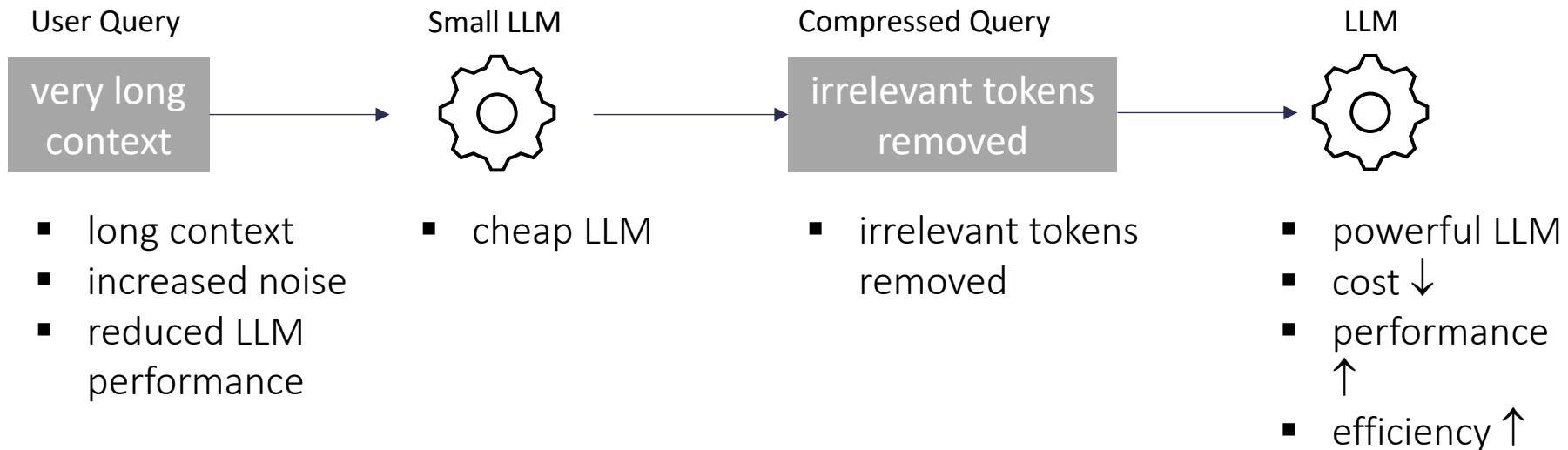


Retrieval-Augmented Generation:

RAG Improvements for
Post-Retrieval

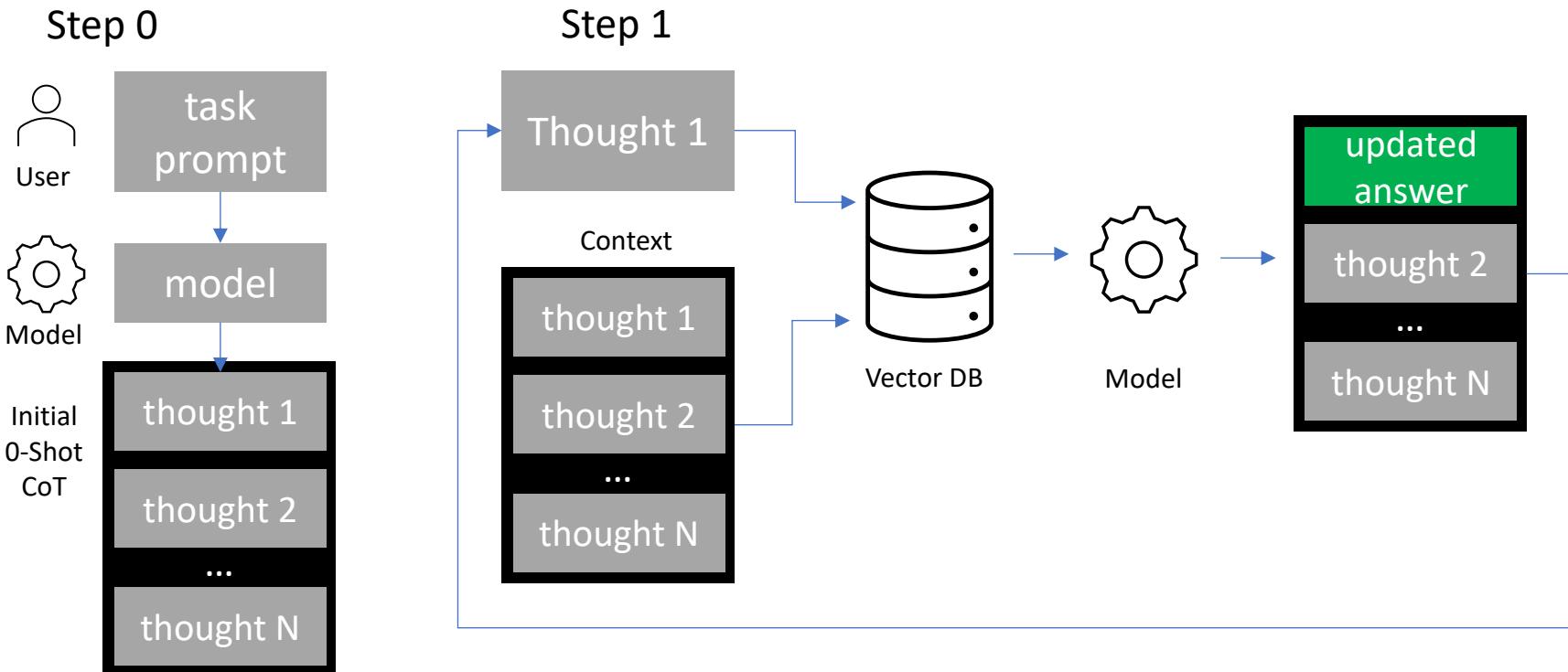
Retrieval-Augmented Generation

RAG Improvement for Post-Retrieval: Prompt Compression



Retrieval-Augmented Generation

Retrieval Augmented Thought



Source: Zihao Wang, et. al. „RAT: Retrieval Augmented Thoughts Elicit Context-Aware Reasoning in Long-Horizon Generation“



gollnickdata.de

Retrieval-Augmented Generation:

RAG Alternatives

Retrieval-Augmented Generation

Prompt Caching

