

Creating a Semantic Search-Based Question-Answering System with LLMs



Jetro Coenradie

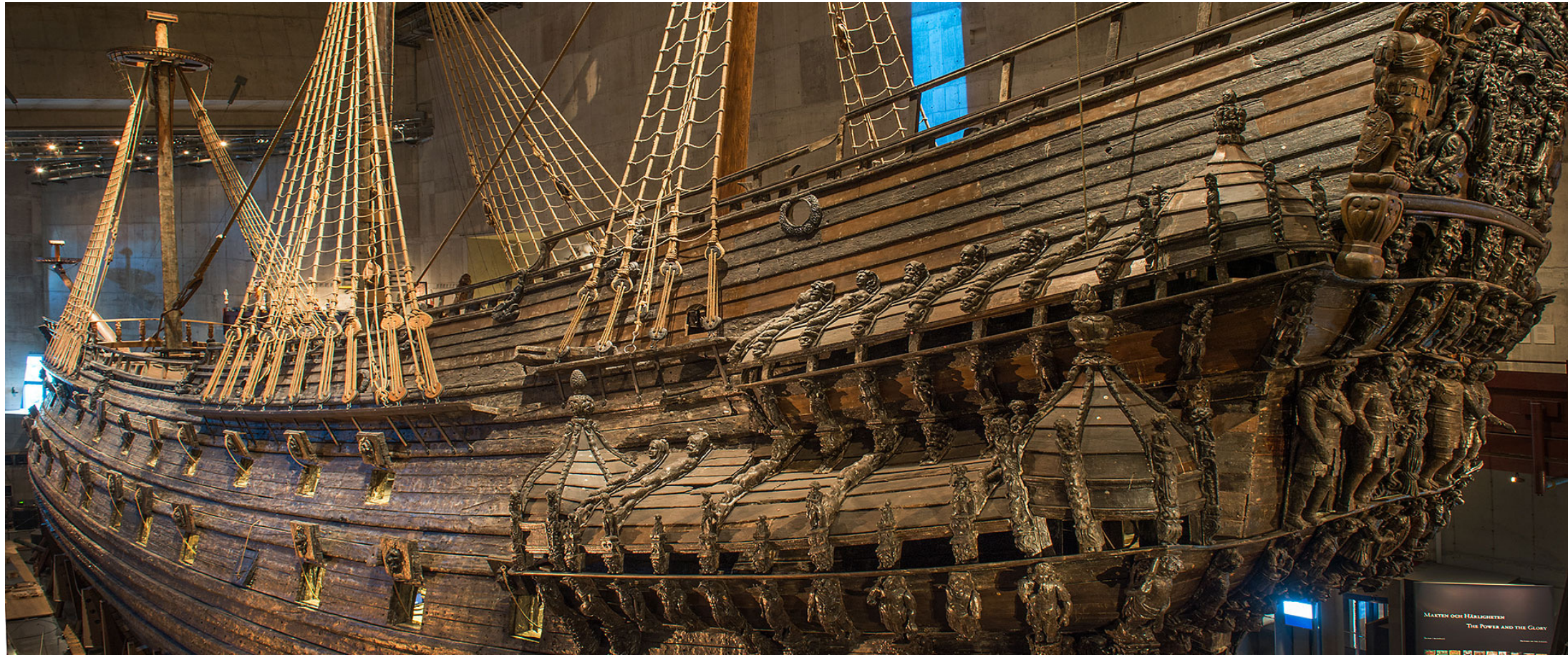


Daniël Spee





The Vasa Warship



<https://www.vasamuseet.se/en>

luminis. / part of
quma

Daniël Spee

Search Engineer @ Luminis

Jfokus24



<https://www.linkedin.com/in/dspee/>



<https://github.com/Daantie>



<https://www.luminis.eu/expert/daniel-spee/>



daniel.spee@luminis.eu

luminis.

Jettro Coenradie

Data Evangelist @ Luminis

Jfokus24



<https://www.linkedin.com/in/jettro>



<https://github.com/jettro>

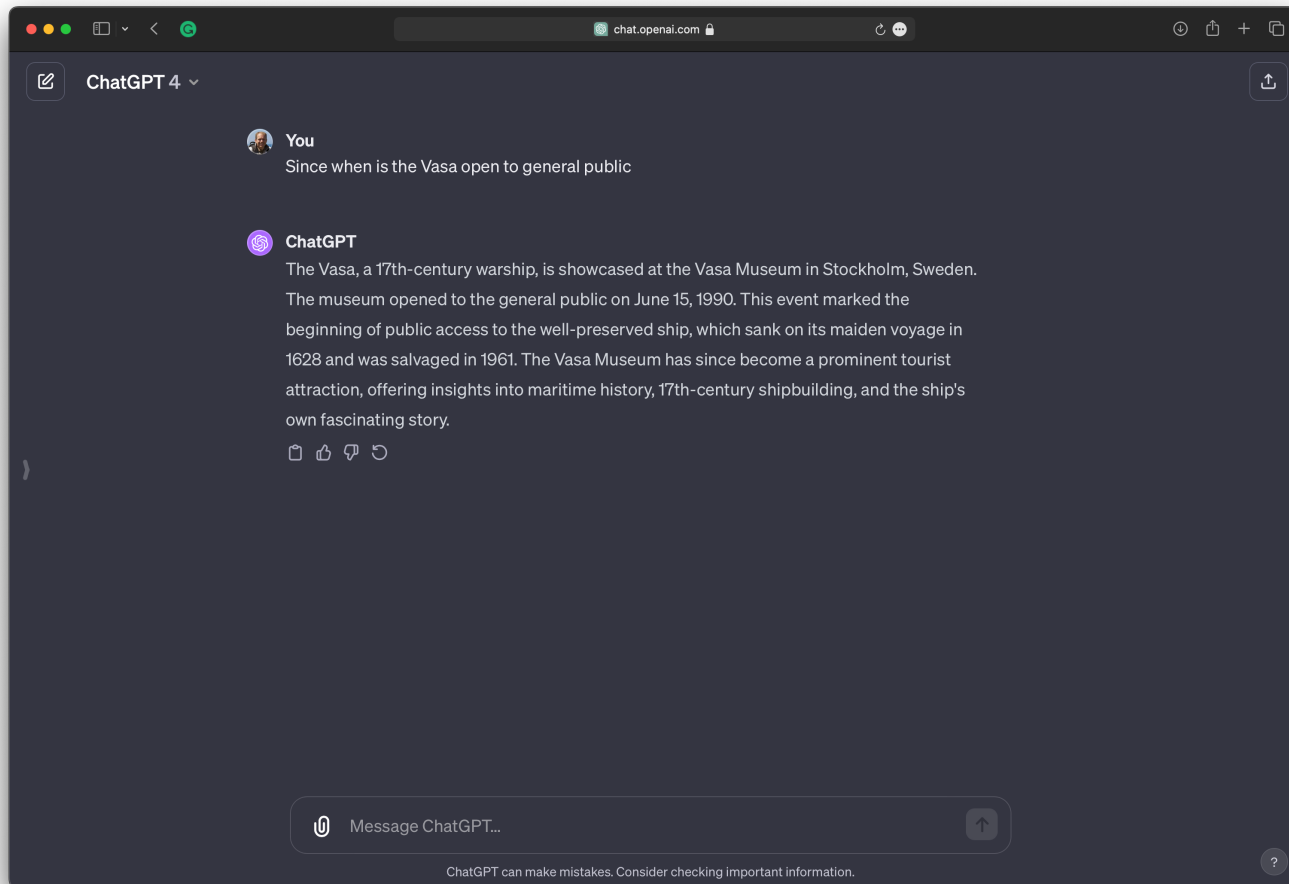


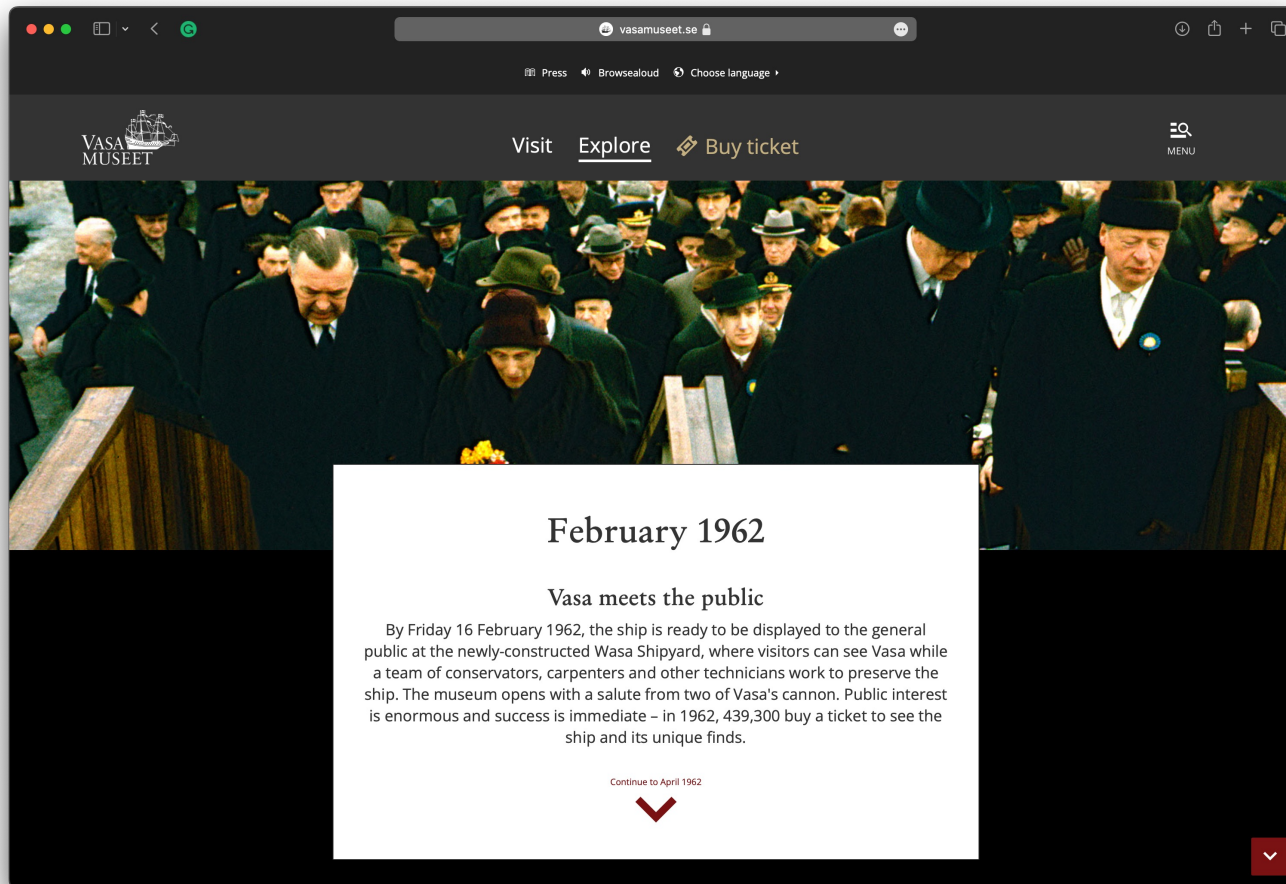
<https://www.luminis.eu/expert/jettro-coenradie/>

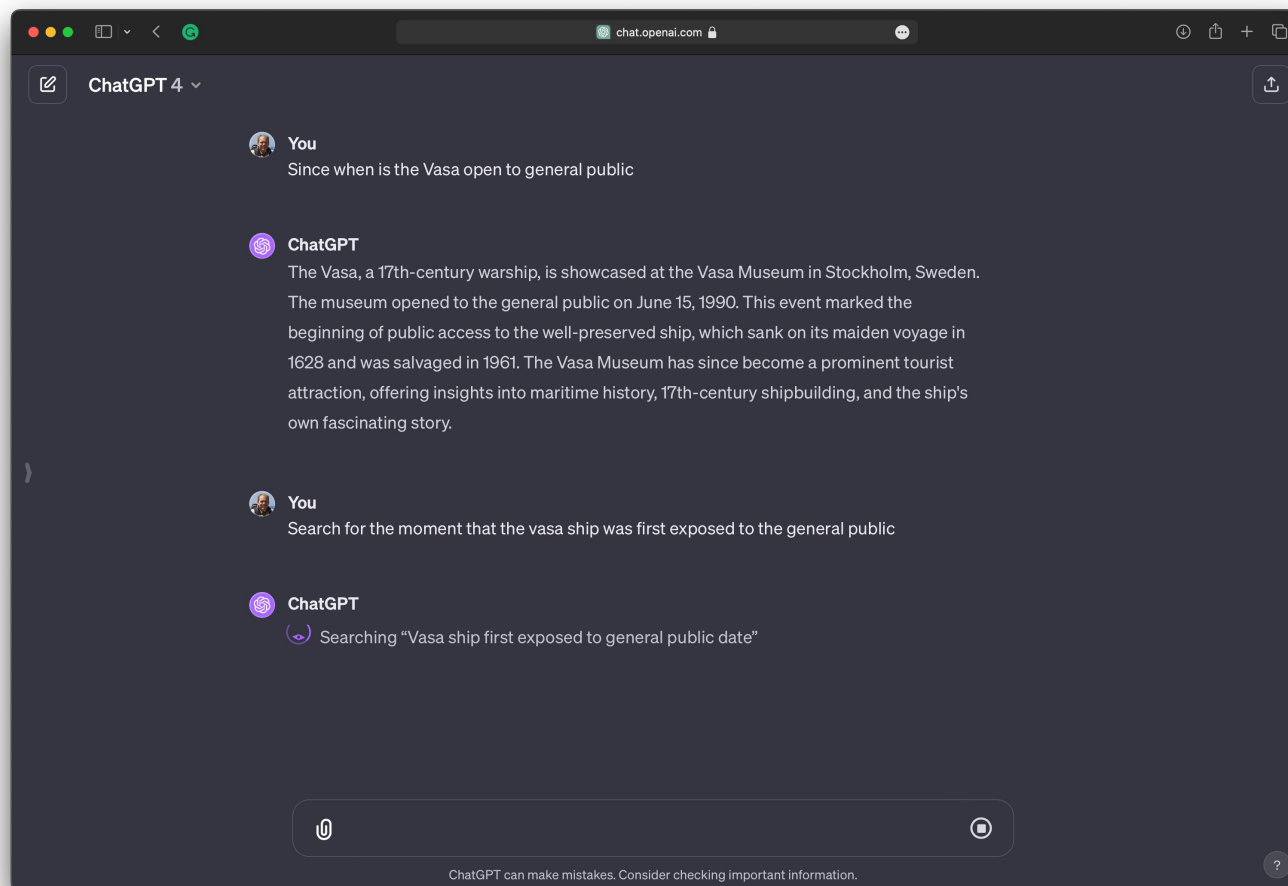


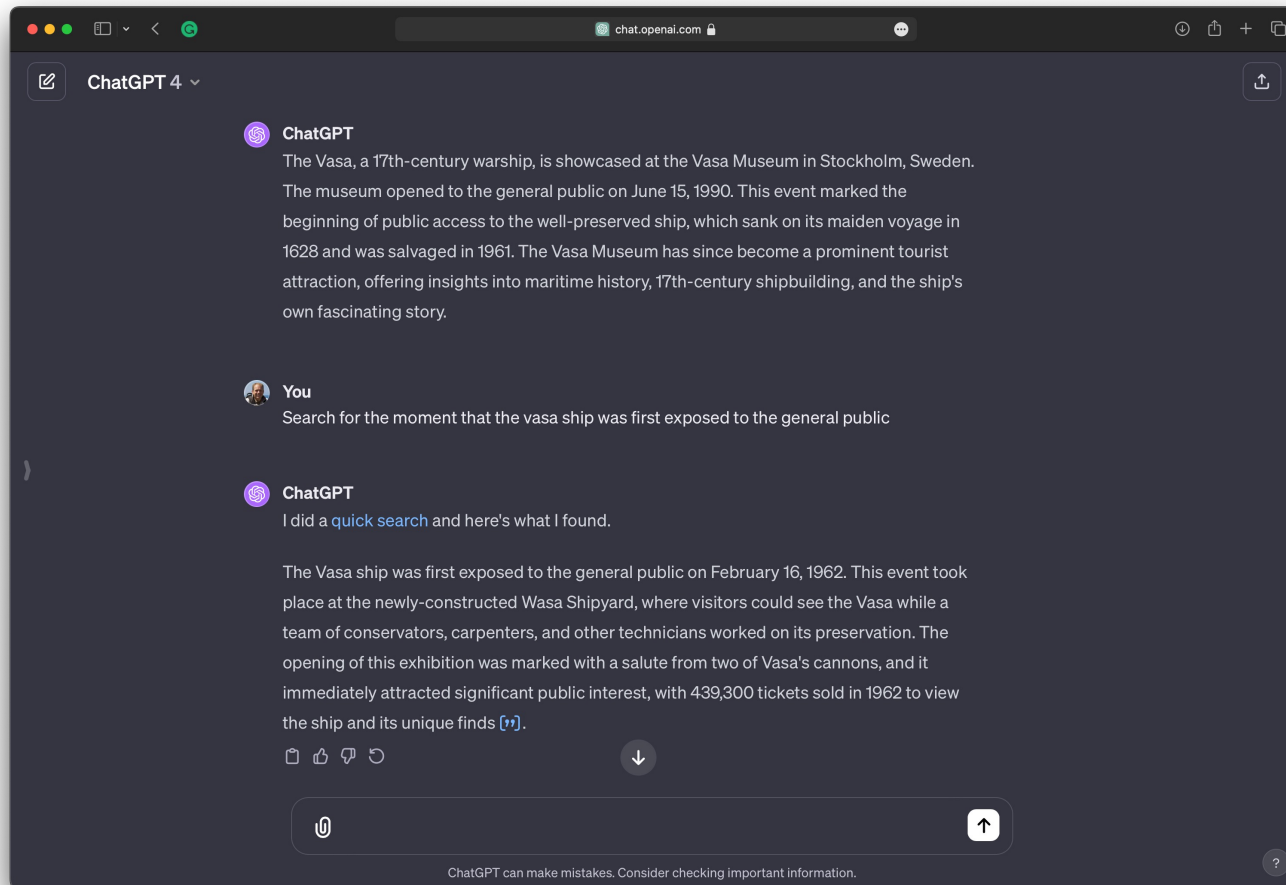
Jettro.coenradie@luminis.eu

luminis.

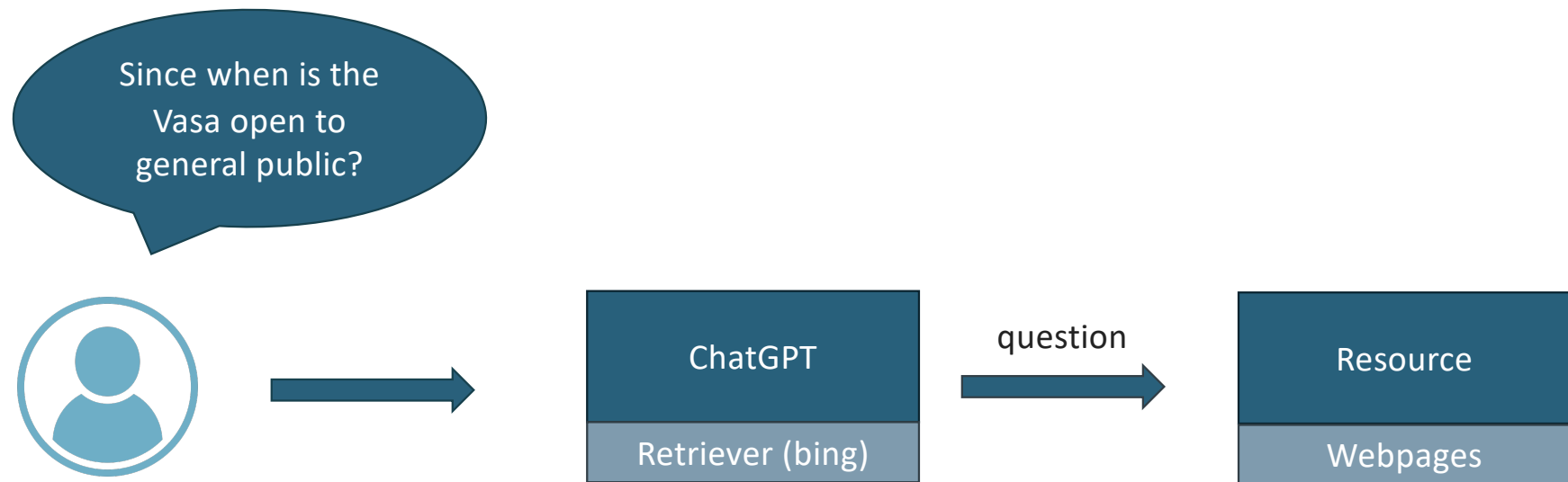




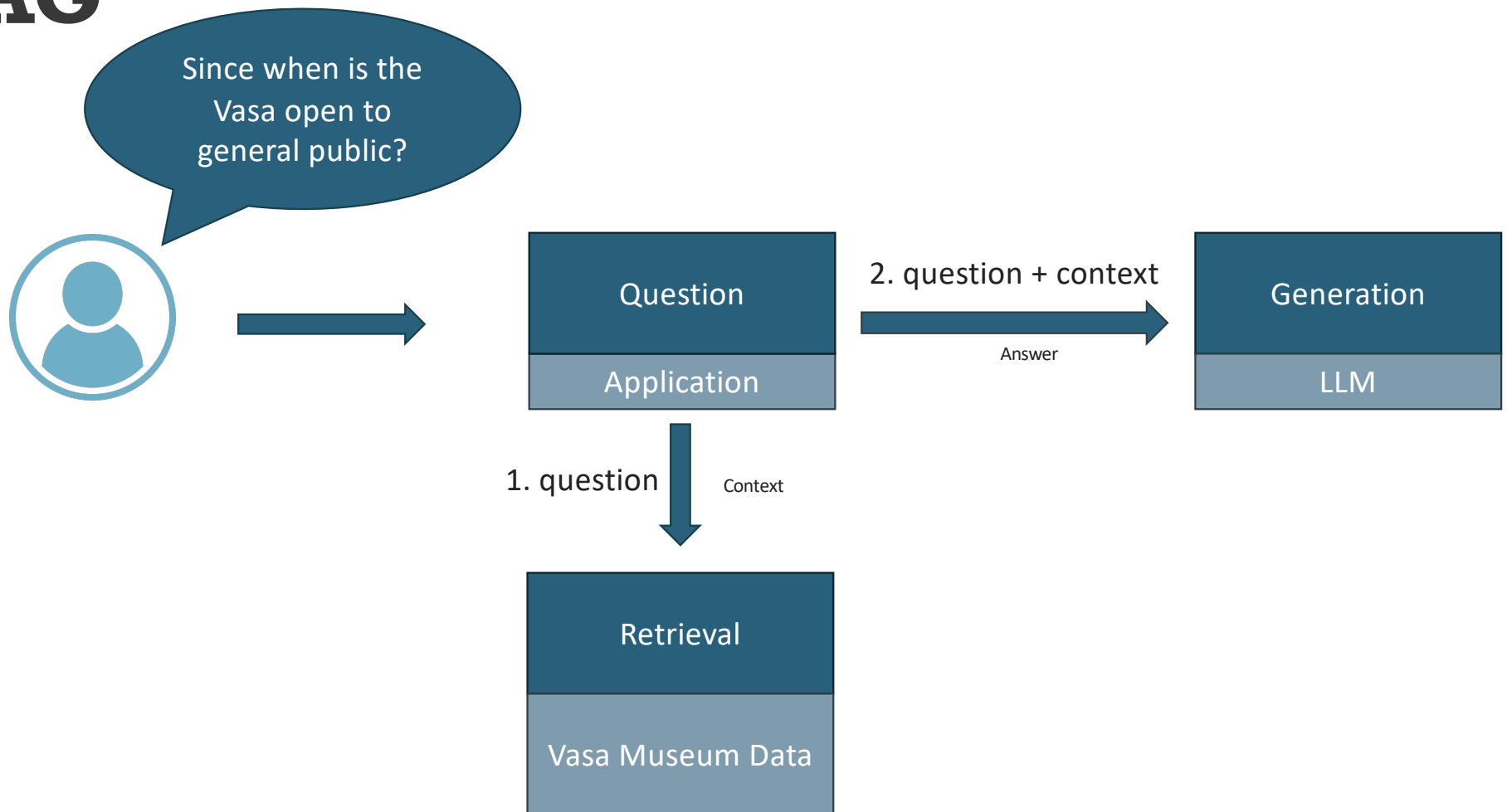




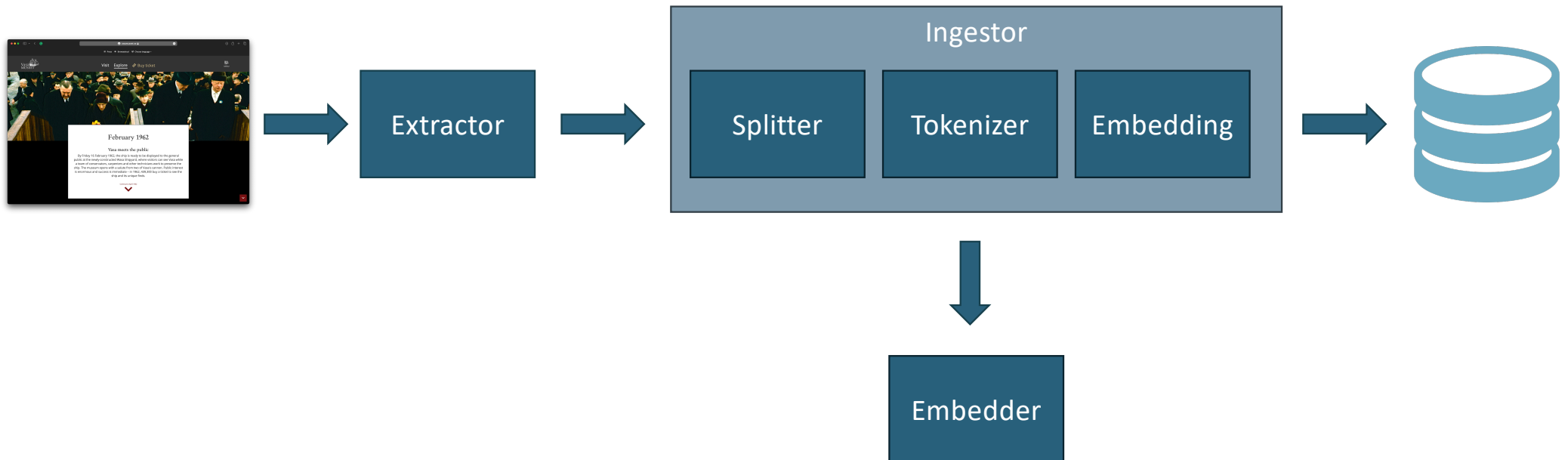
Retrieval Augmented Generation (RAG)



RAG



Vector Store ~ Ingestion



Workshop agenda



Setting up your environment



Splitting text into chunks



Embedding chunks



Storing embedding and data in a content store

Use synthetic questions to validate retrieval quality.



Generate an answer to a question from a context

Use LLM to validate the quality of the answer



Improve retrieval strategy



All together now

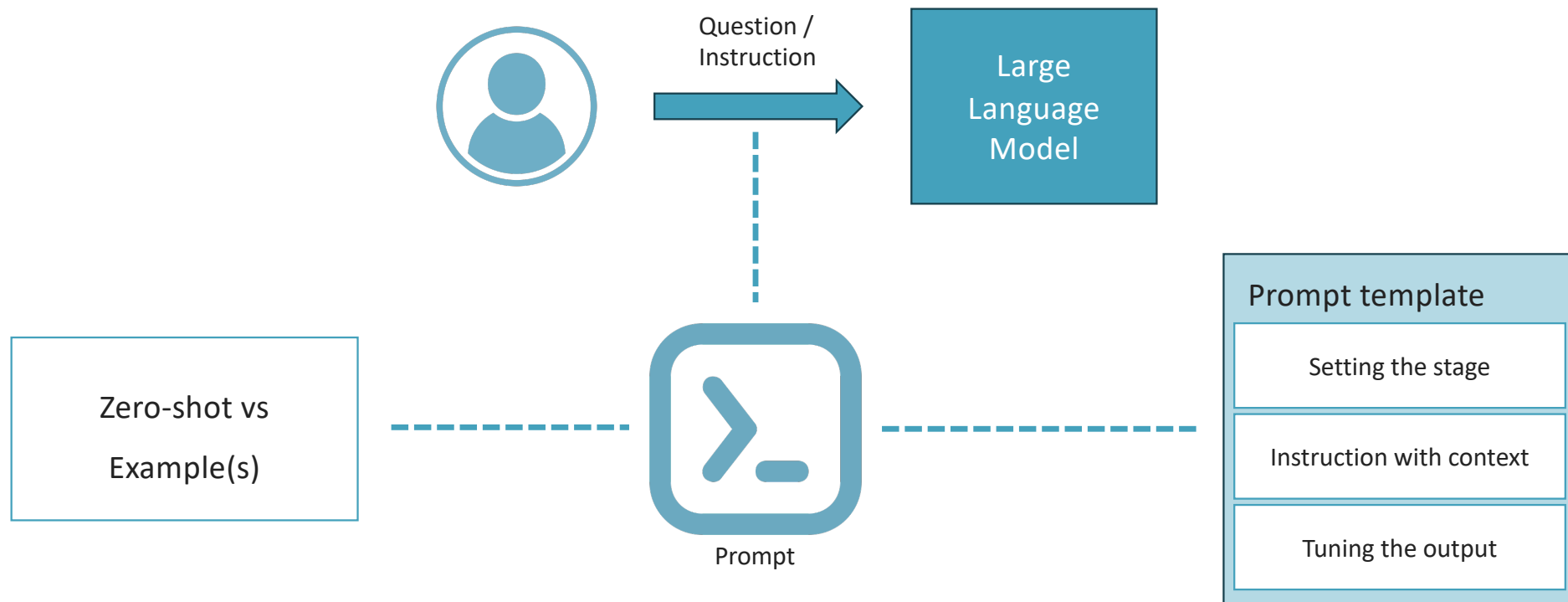
LLM – Large Language Model



Guardrules / Instructions / Parent prompting



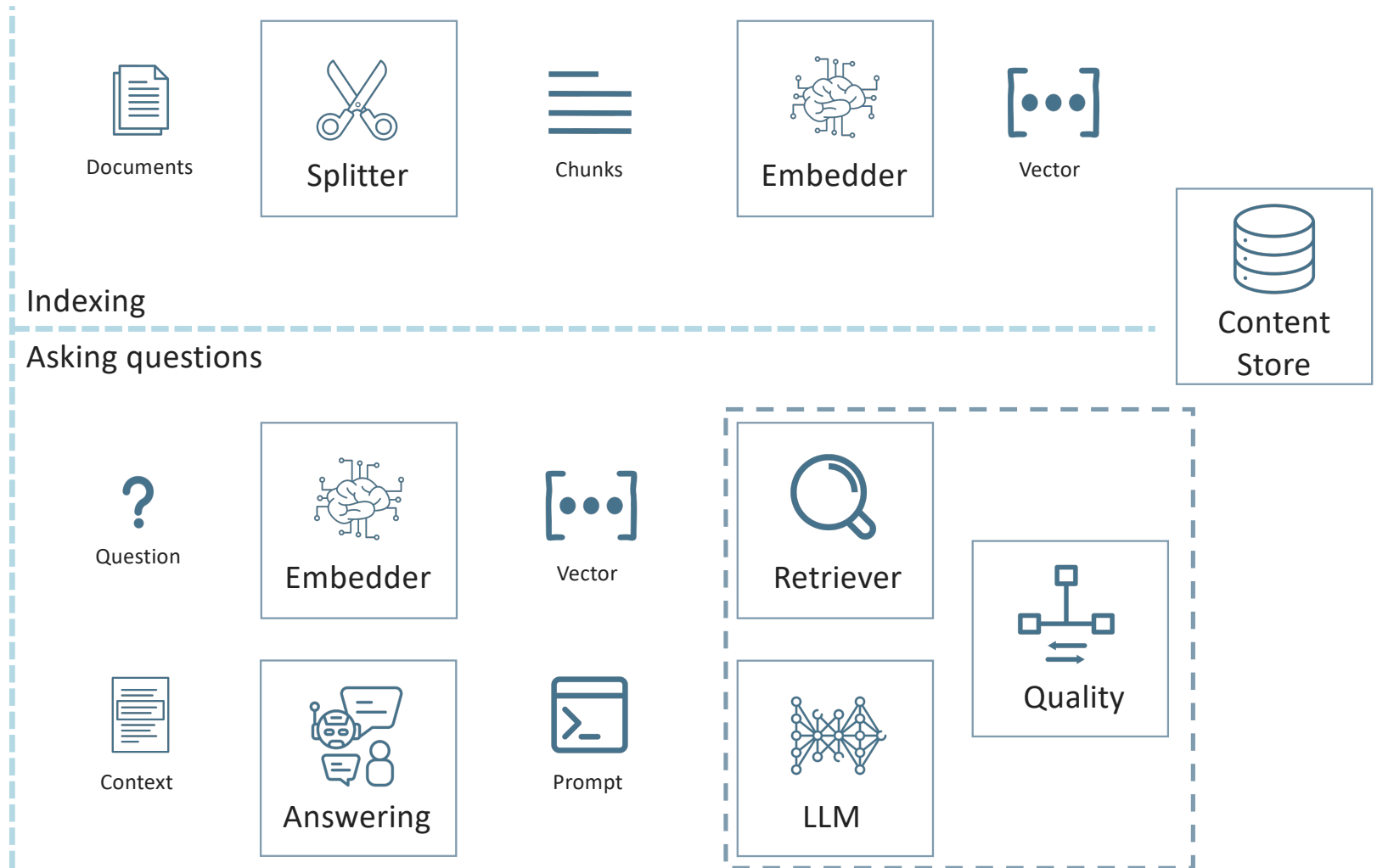
Prompting





Java: AppStepN...

Python: app_stepN_...



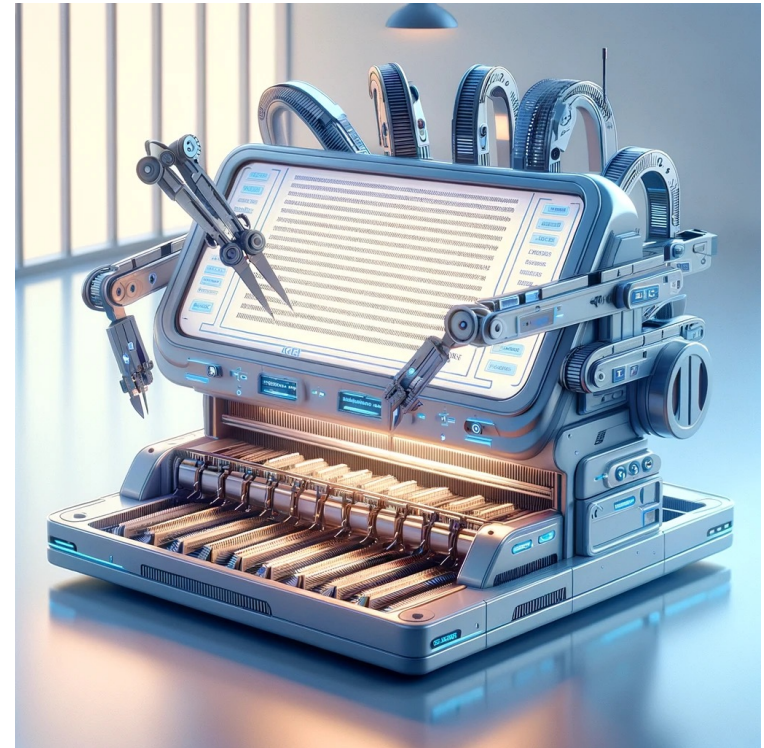
Chunking Strategy

Splitting text into chunks

Why is splitting necessary?

Impact of different splitters

Examples splitters

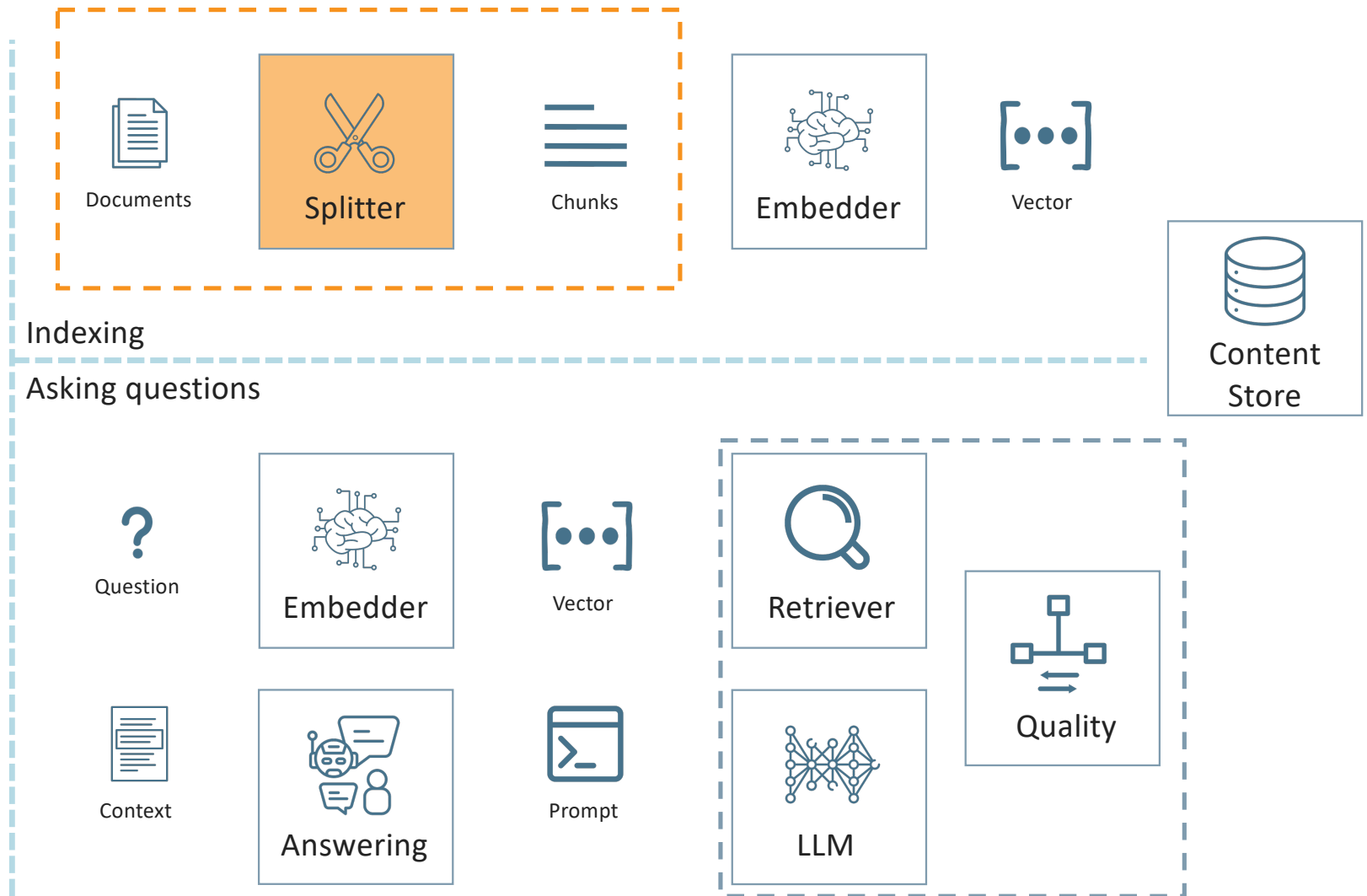


DALL E - visualize a splitter that splits text into chunks



Java: AppStepN...

Python: app_stepN_...





<https://rag4j.org>

Workshop jFokus 2024

<https://github.com/RAG4J/rag4j-jfokus>

<https://github.com/RAG4J/rag4p-jfokus>

secret_key=

J: AppStep1ChunkingStrategy
P: app_step1_chunking_strategy

Embeddings

Vectorisation

From text to numbers

- Dictionaries of tokens
- Sparse vectors
- One hot encoding

Dictionary

programming	1	0	0	0	0
search	0	1	0	0	0
data	0	0	1	0	0
language	0	0	0	1	0
pipeline	0	0	0	0	1

Document

programming	1	0	0	0	0
your	0	0	0	0	0
data	0	0	1	0	0
Pipeline	0	0	0	0	1

Query

programming	1	0	0	0	0
language	0	0	0	1	0

Dictionaries

From text to numbers

- Tokenisation
- Wordparts
- GPT-4: 100277

English

'From'	'legacy'	'to'	'the'	'cloud'	':'	'how'	'modern'	'software'	'development'	'speeds'	'you'
3915	20160	311	279	9624	25	1268	6617	3241	4500	25753	499
'up'	'.'										
709	627										

Nederlands

'E'	'en'	'hy'	'br'	'ide'	'zoek'	'op'	'loss'	'ing'	'in'	'plaats'	'van'	'SA'	'AS'	':'
36	268	6409	1347	579	74900	454	9563	287	304	74957	5355	16998	1950	25
'eff'	'ici'	'ë'	'nt'	'er'	'en'	'met'	'un'	'ie'	'ke'	'karakter'	'ist'	'ie'	'ken'	'.'
3369	3457	12456	406	261	665	2322	653	648	441	91855	380	648	2779	627

From vectors to concepts

How are vectors or embeddings created

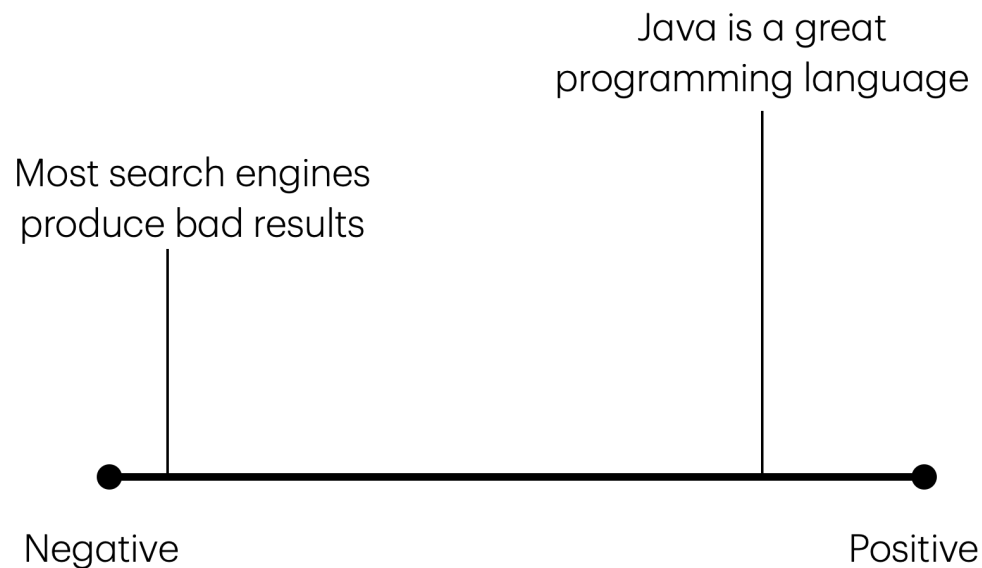
'Java'	'is'	'a'	'great'	'programming'	'language'	' '
15391	374	264	2294	15840	4221	198

Embedding - [0.85]

'Most'	'search'	'engines'	'produce'	'bad'	'results'	' '
13622	2778	21787	8356	3958	3135	198

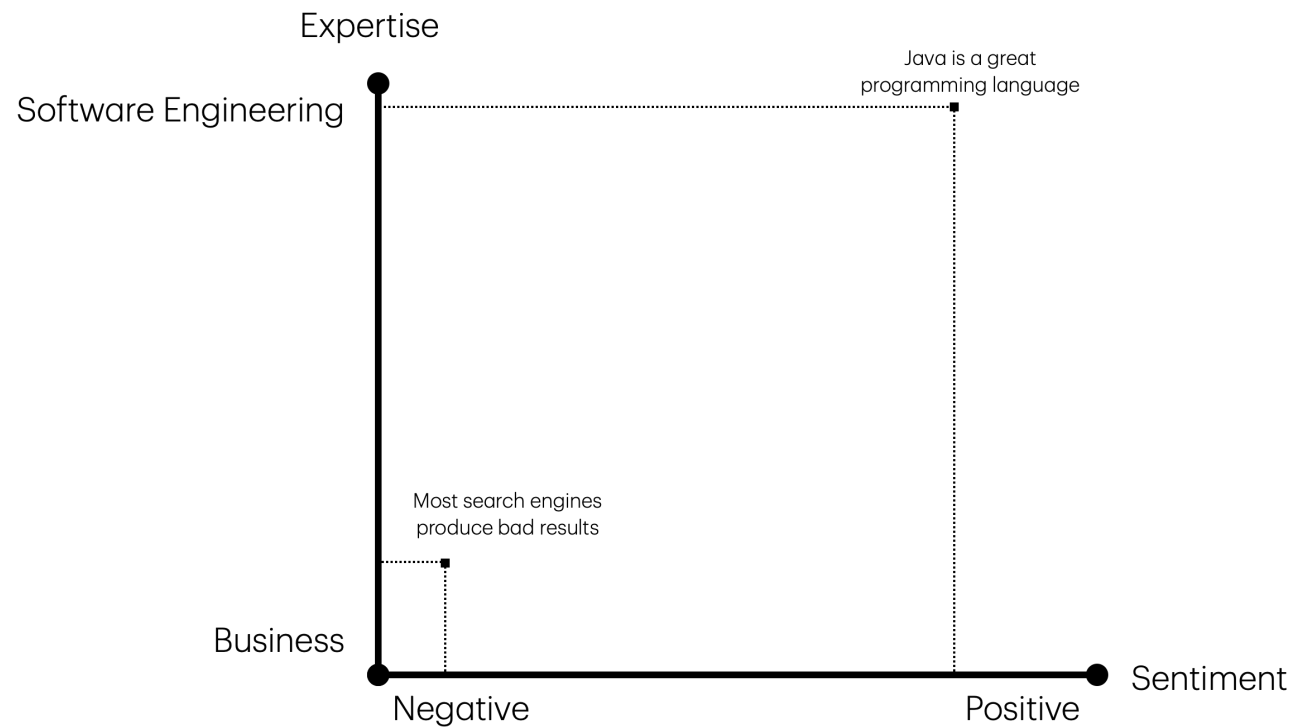
Embedding - [0.09]

Sentiment



From vectors to concepts

How are vectors or embeddings created



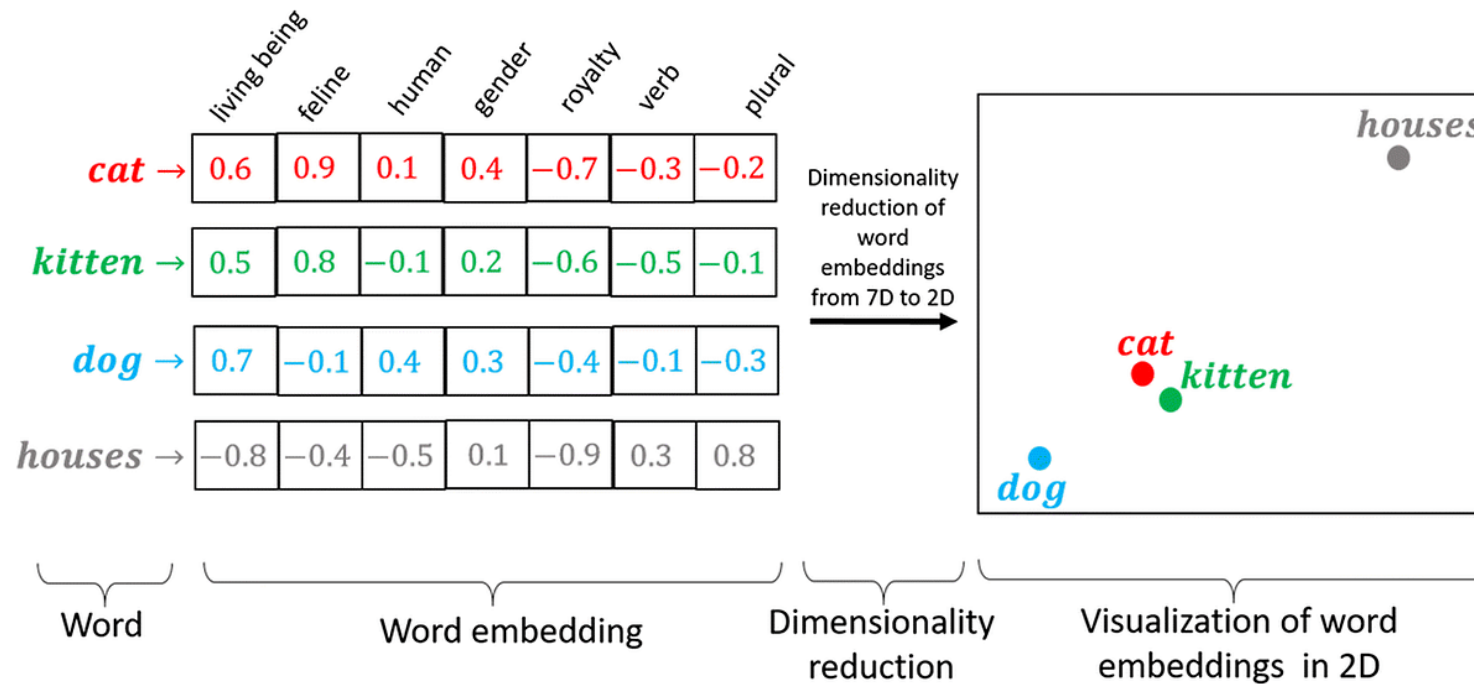
'Java'	'is'	'a'	'great'	'programming'	'language'	' '
15391	374	264	2294	15840	4221	198

Embedding - [0.85, 1]

'Most'	'search'	'engines'	'produce'	'bad'	'results'	' '
13622	2778	21787	8356	3958	3135	198

Embedding - [0.09, 1.3]

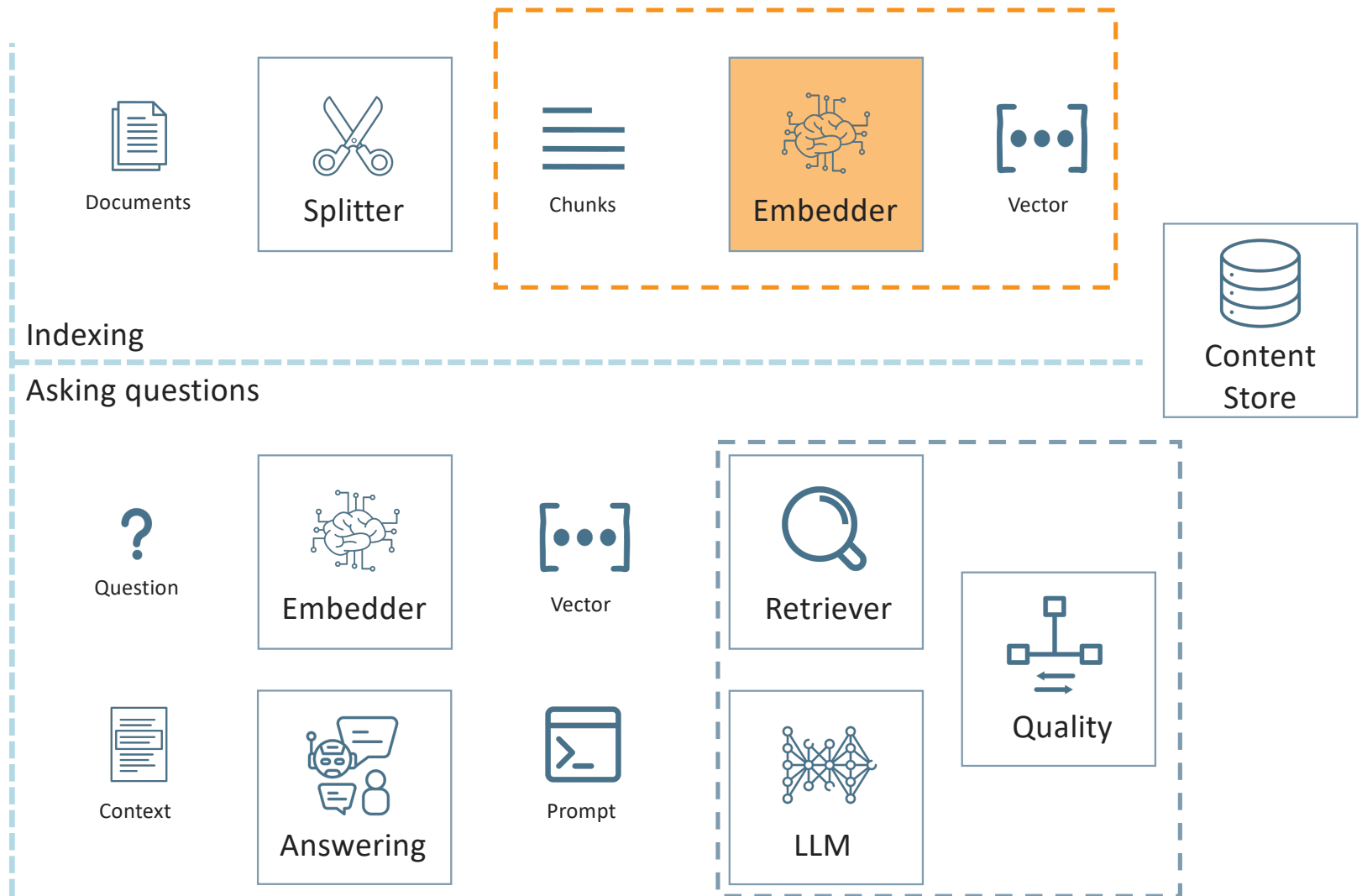
Embeddings





Java: AppStepN...

Python: app_stepN_...





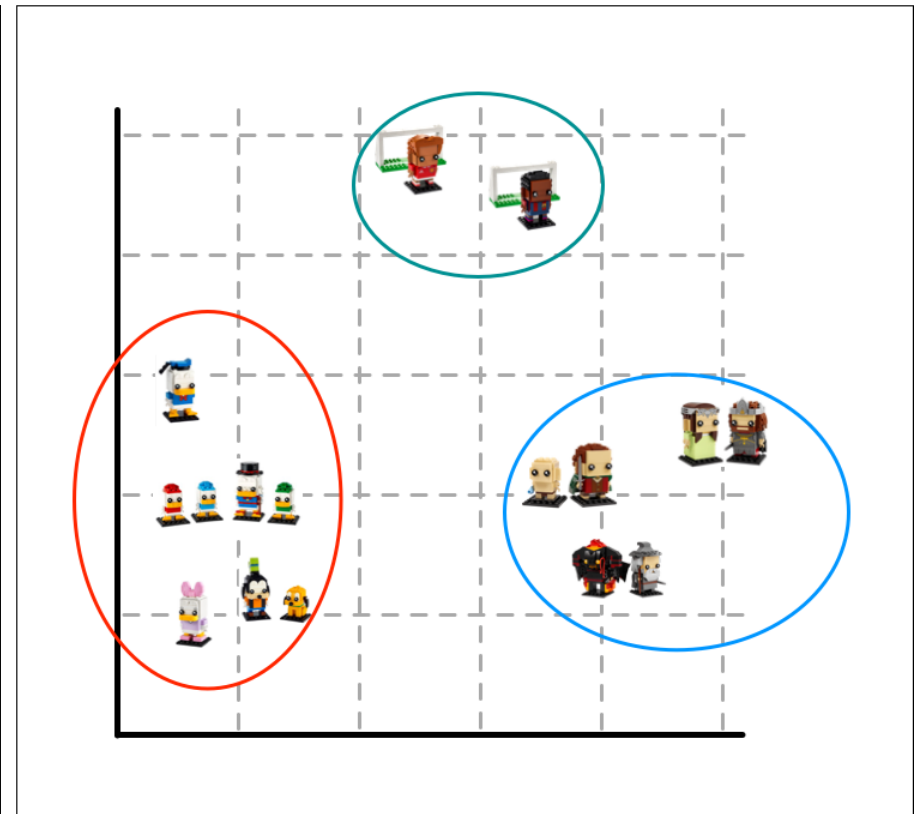
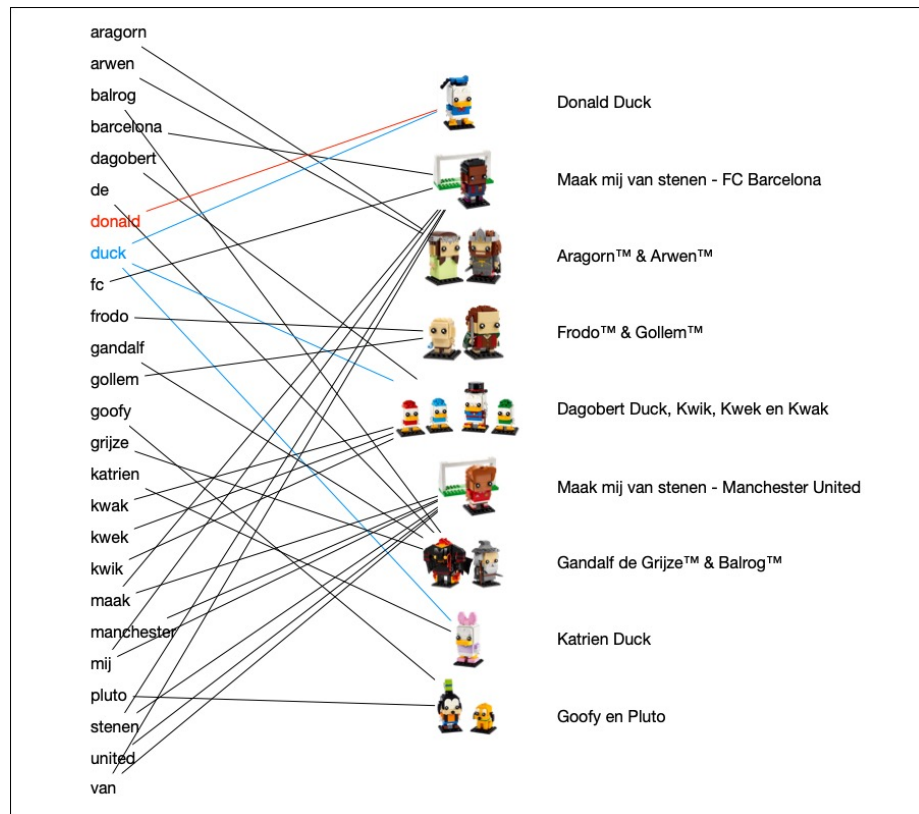
secret_key=

J: AppStep2Embedder

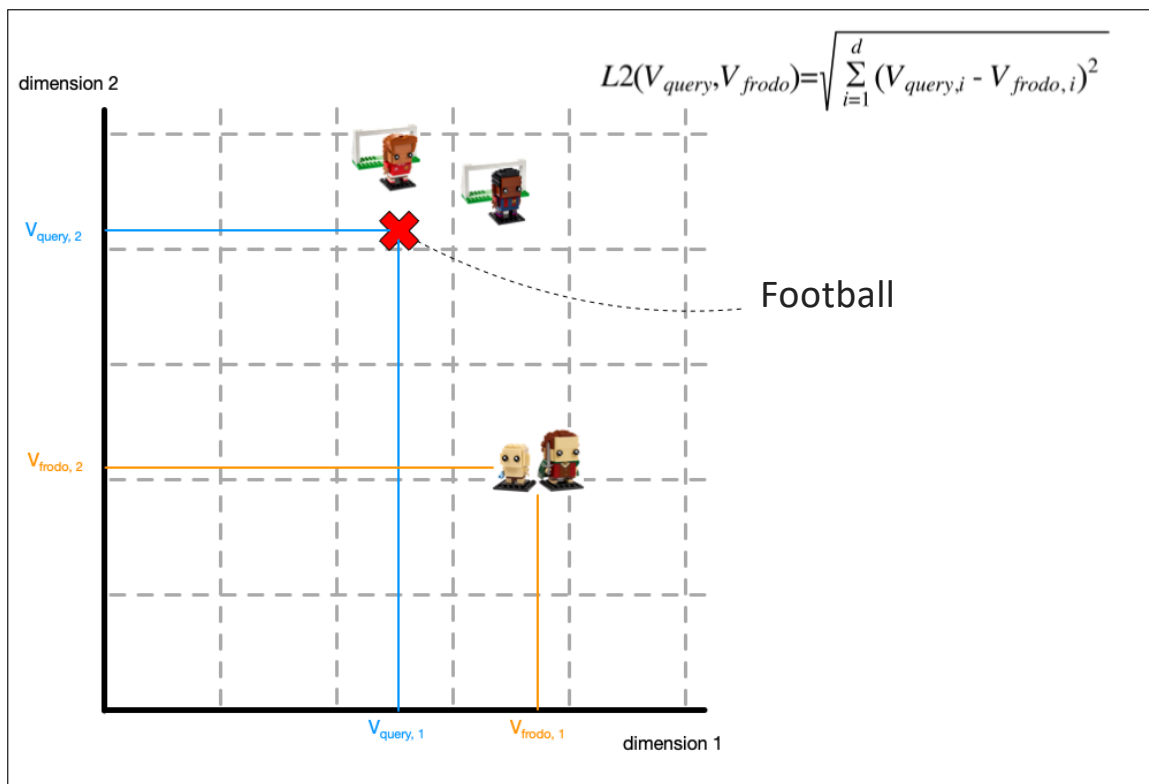
P: app_step2_embedder

Content Stores

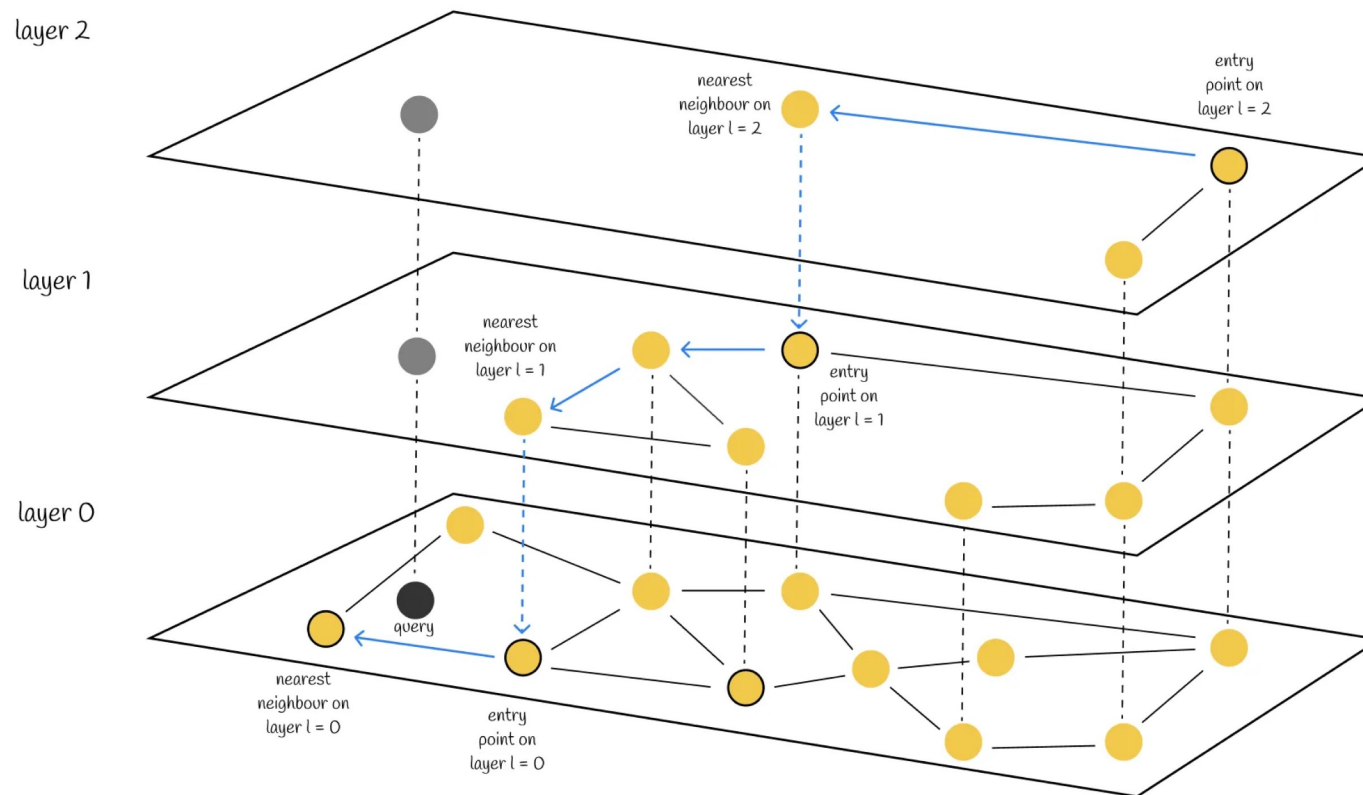
Lexical vs Semantic Search



Comparing vectors



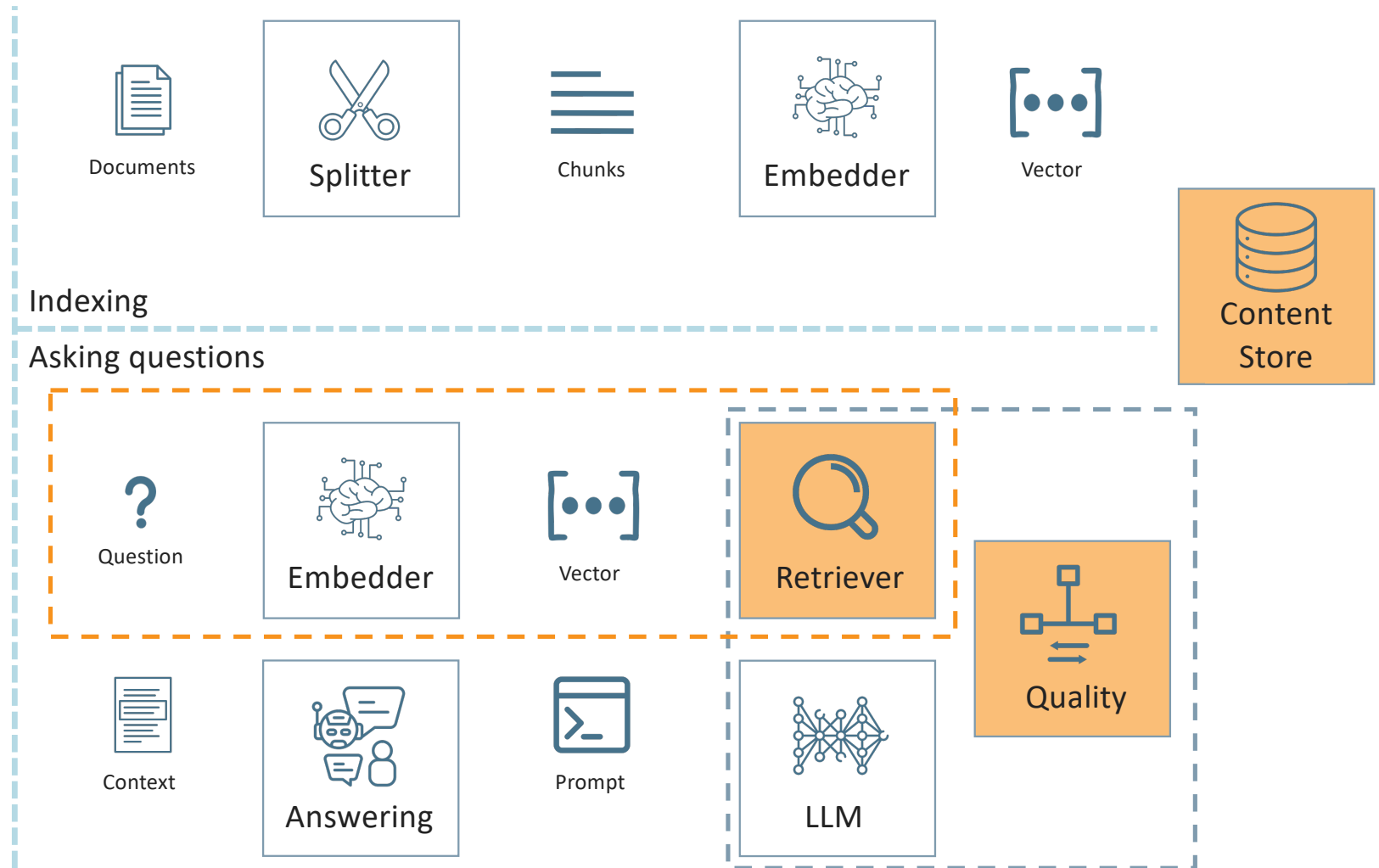
Similarity / HNSW





Java: AppStepN...

Python: app_stepN_...





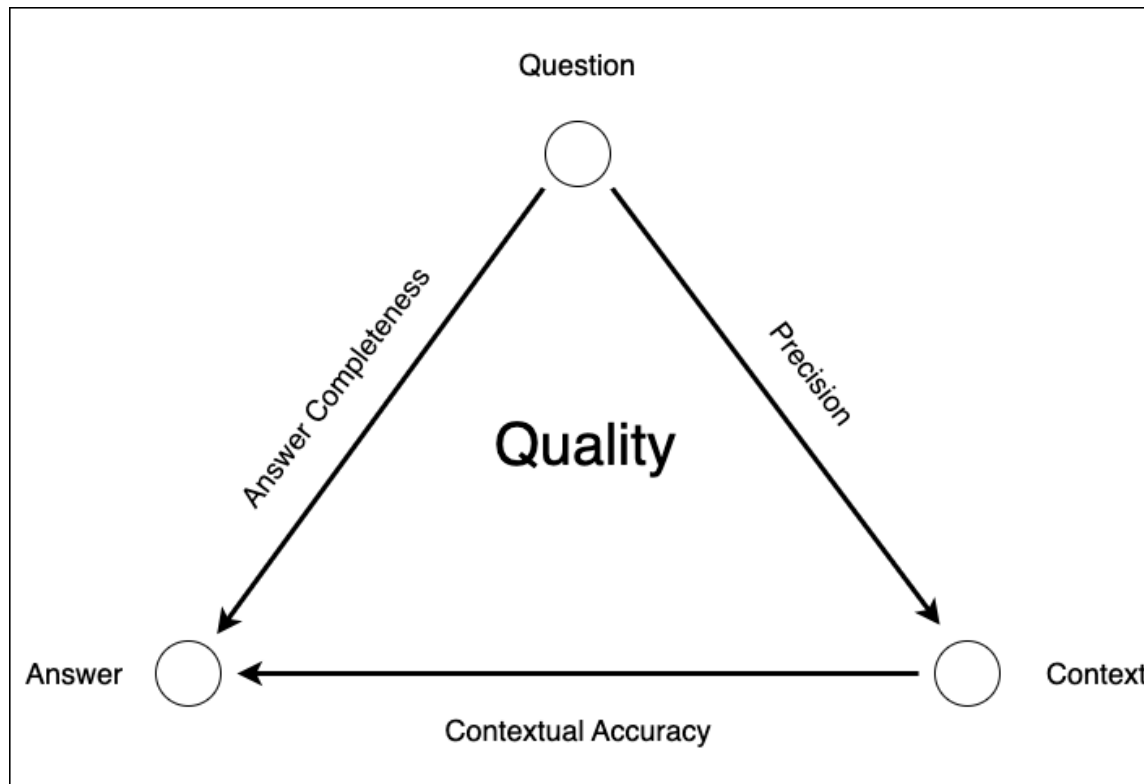
secret_key=

J: AppStep3ContentStore

P: app_step3_content_store

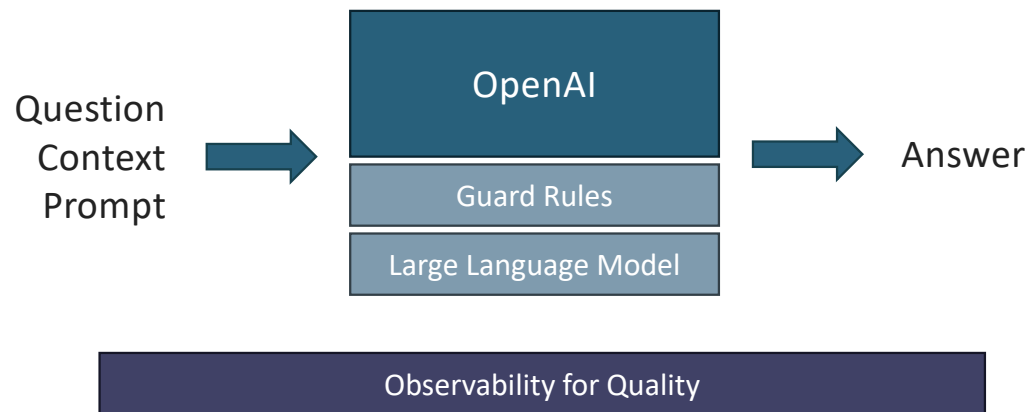
Evaluation

Quality of RAG



Answer Generation

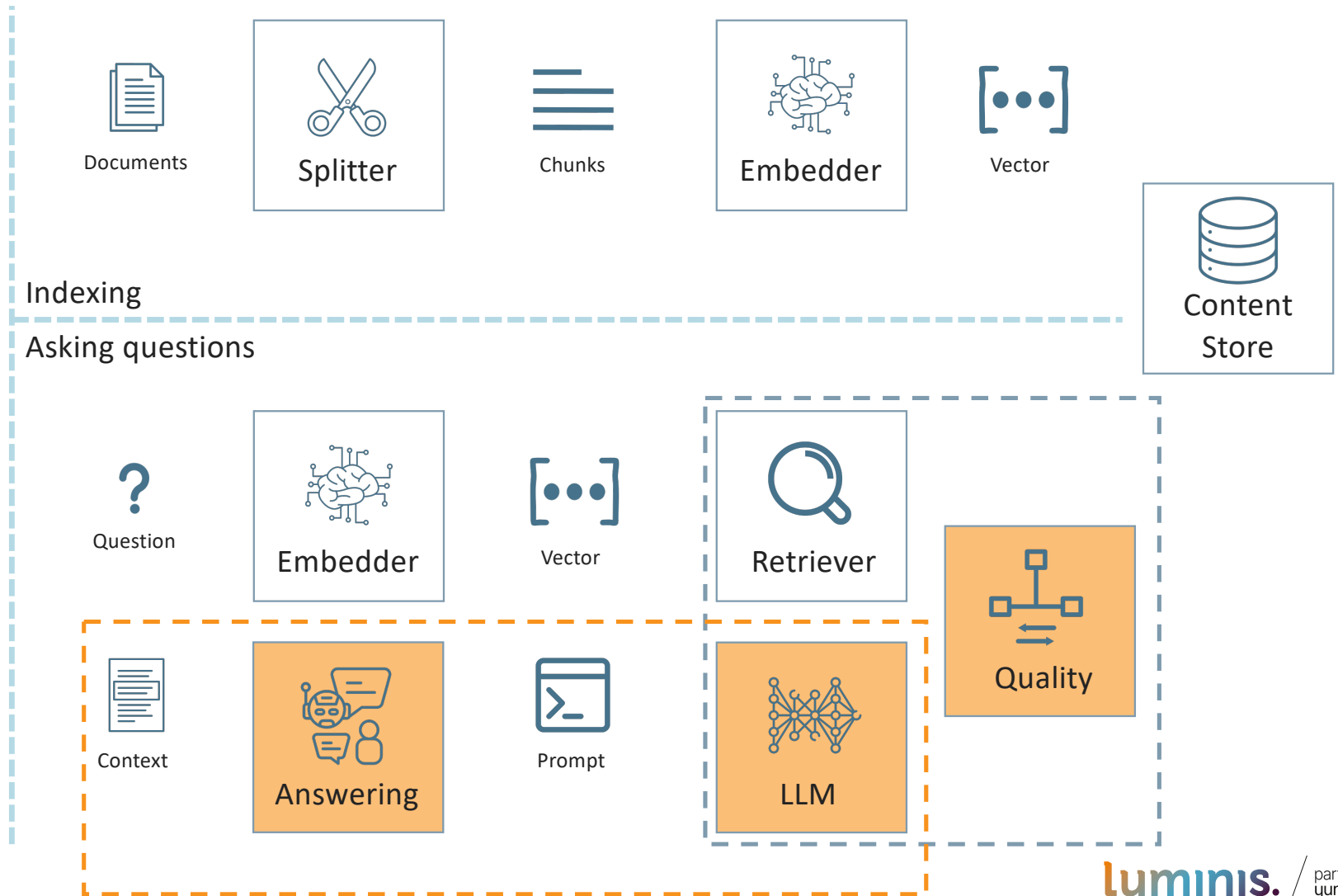
Generating Answers





Java: AppStepN...

Python: app_stepN_...





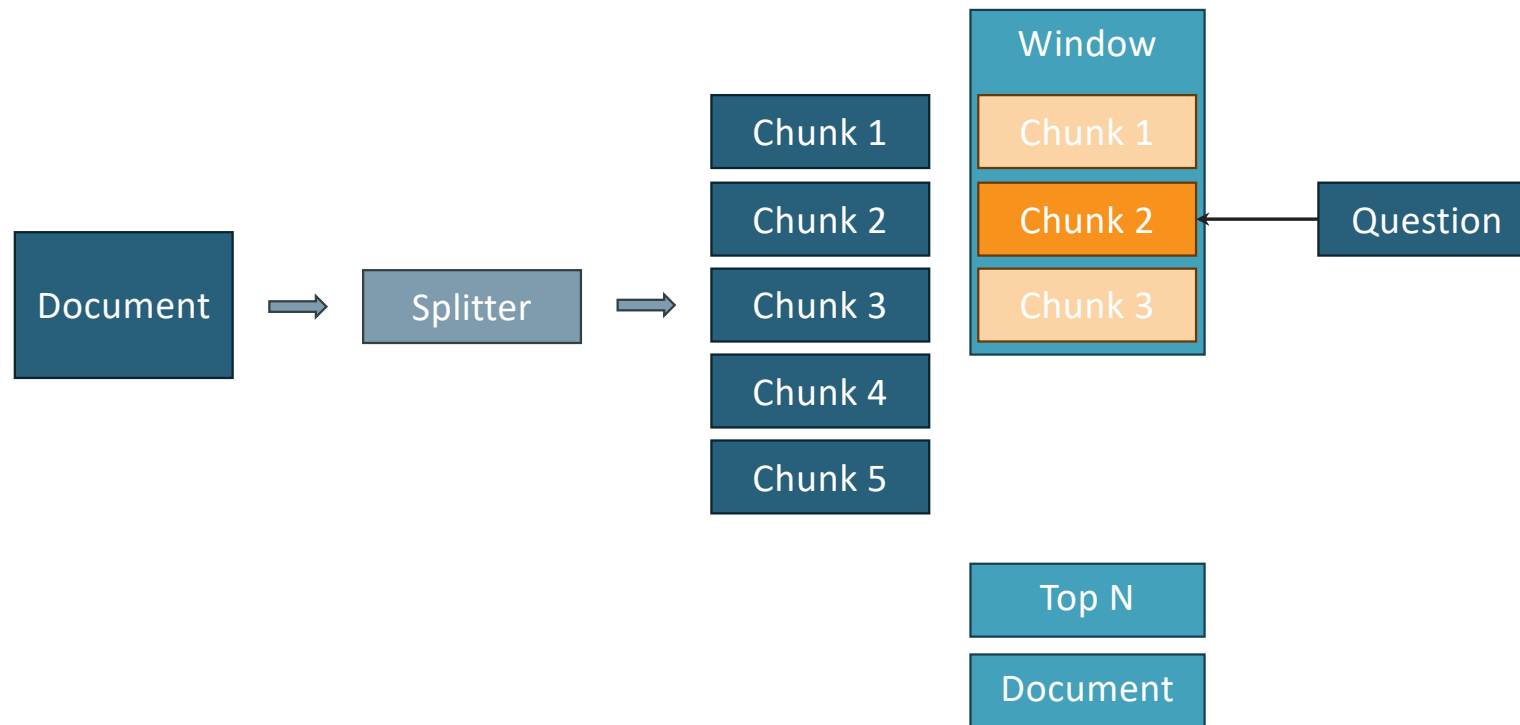
secret_key=

J: AppStep4Generator

P: app_step4_generator

Retrieval Strategy

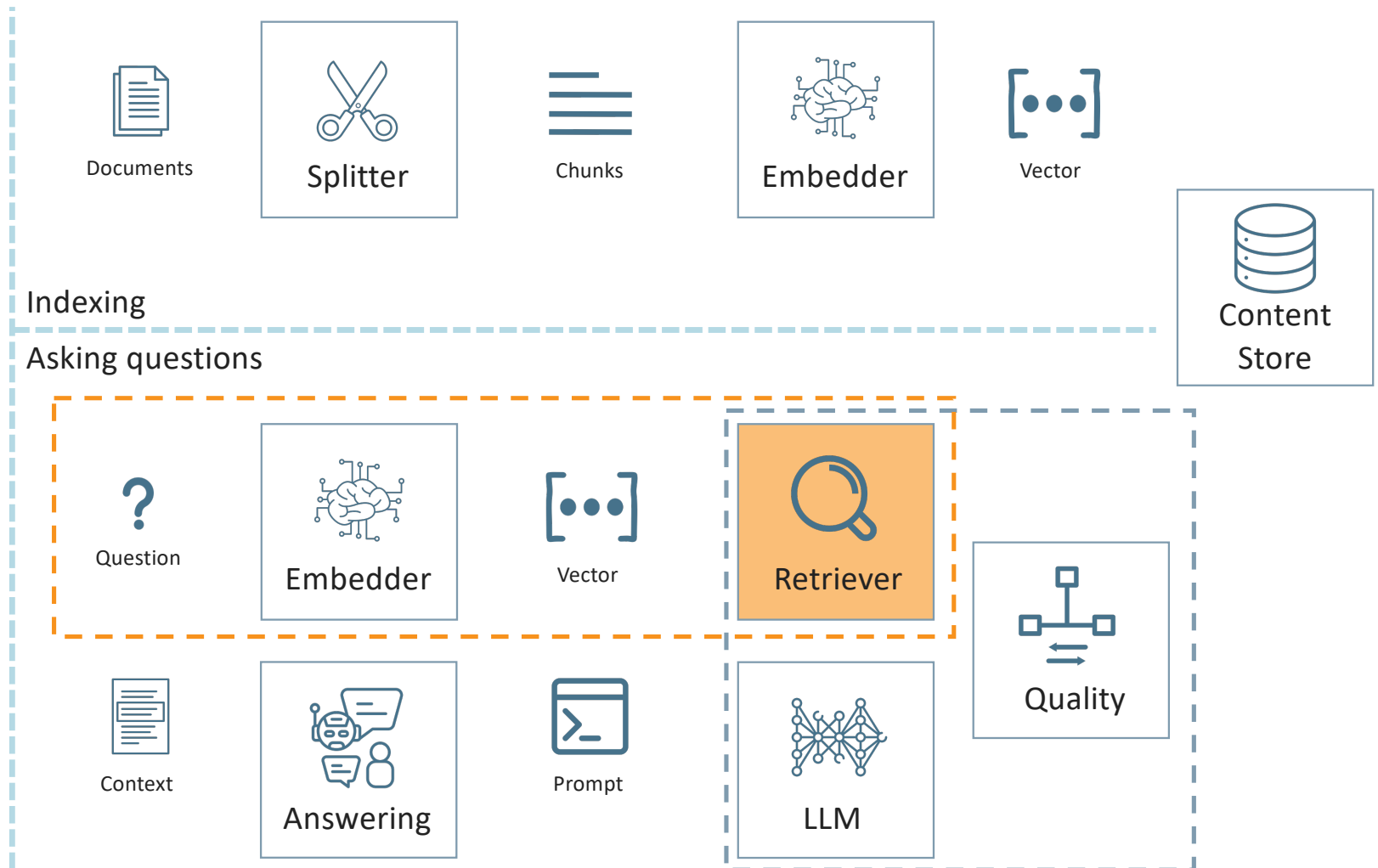
Retrieving the context





Java: AppStepN...

Python: app_stepN_...



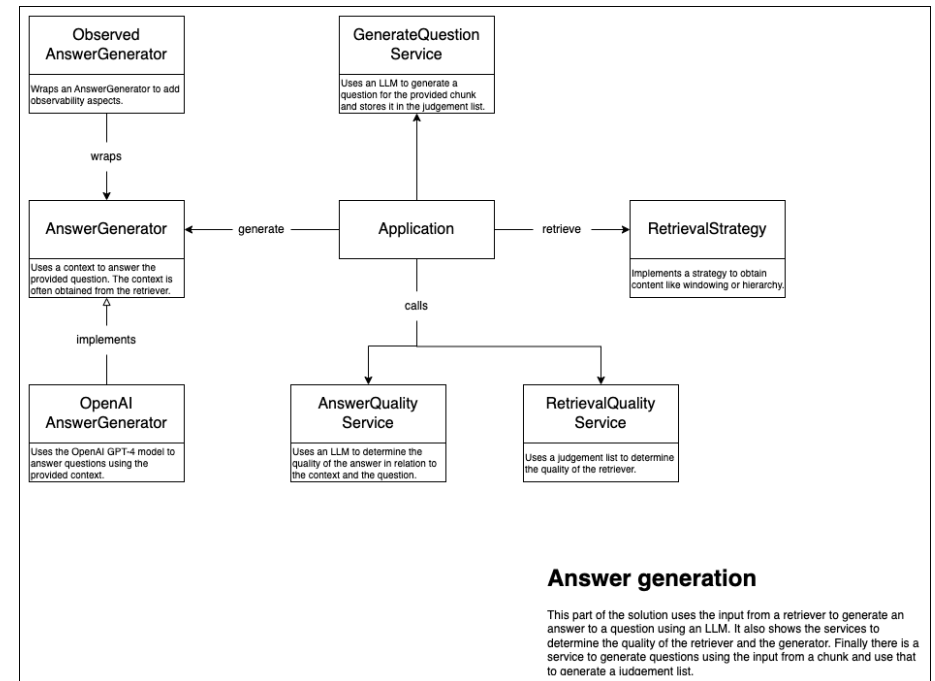
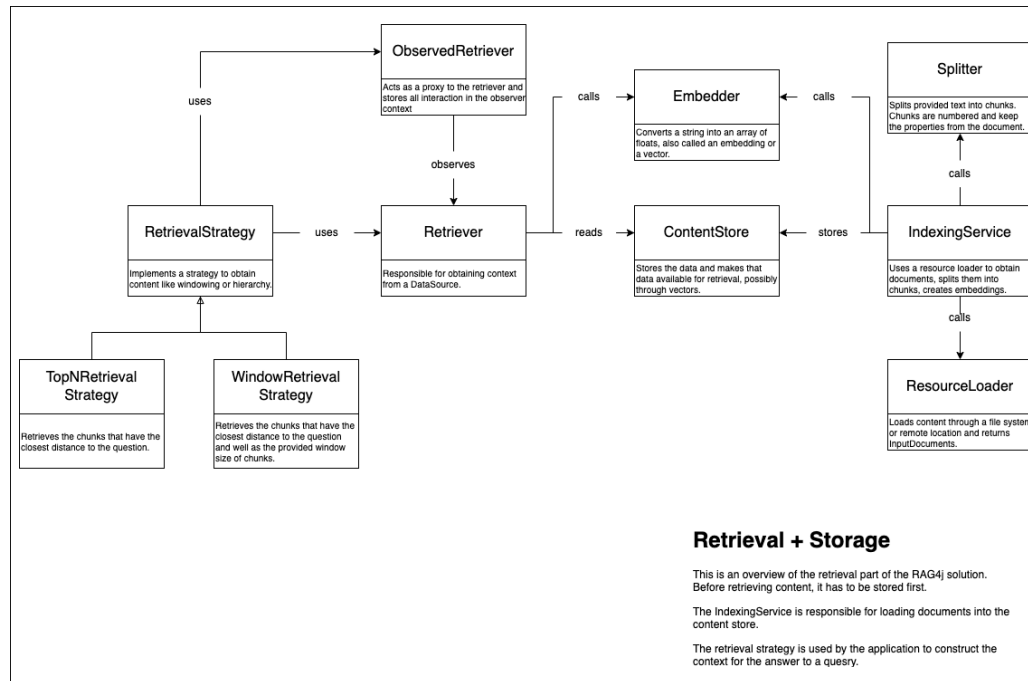


secret_key=

J: AppStep5RetrievalStrategy
P: app_step5_retrieval_strategy

Everything together

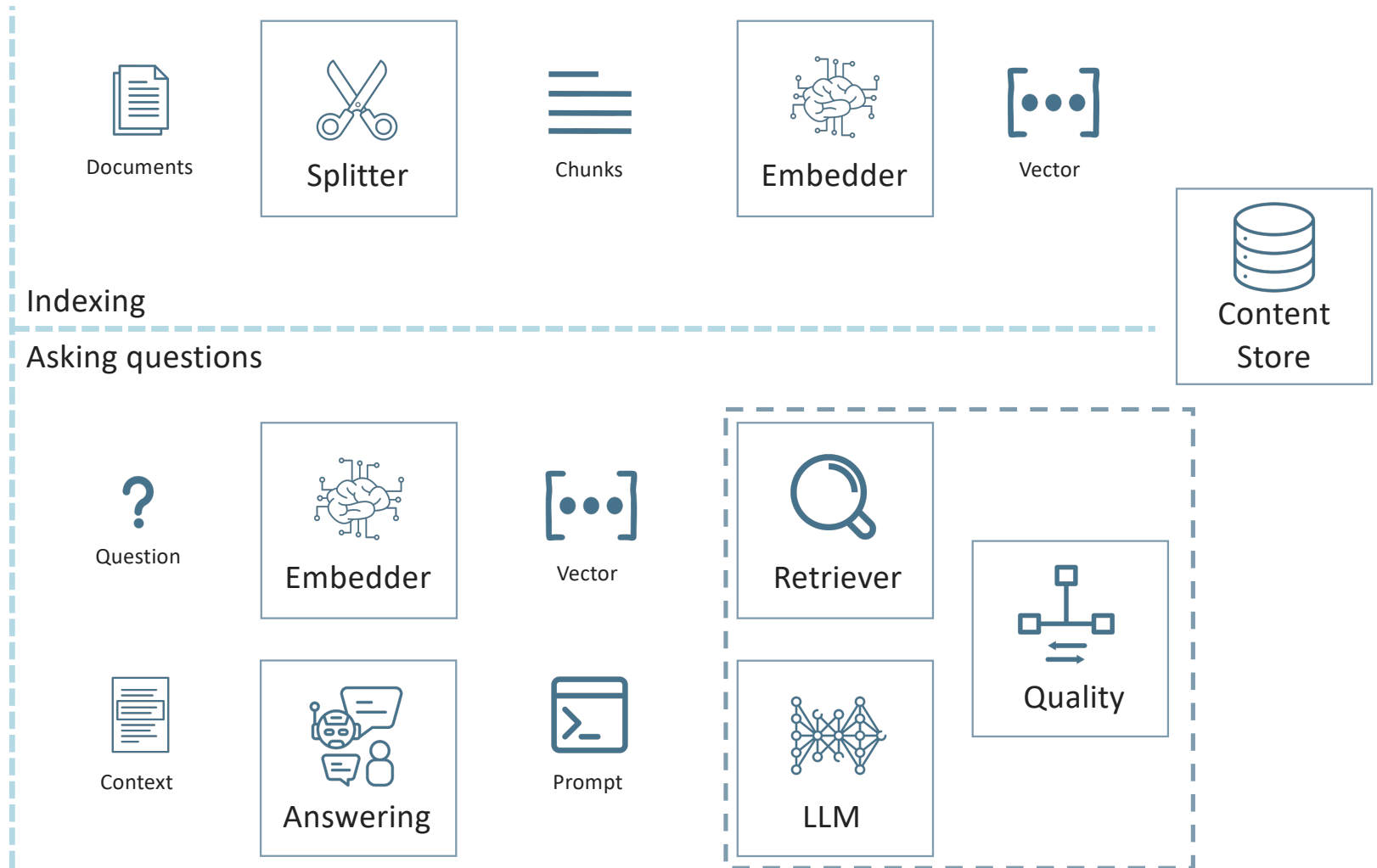
Overall Solution





Java: AppStepN...

Python: app_stepN_...





secret_key=

J: AppStep6AllTogether

P: app_step6_all_together

Thanks!



Daniël Spee



Jetro Coenradie