Master Seminar

on

Development of a Knowledge Retrieval Chatbot By integrating Elastic Search Based Retrieval Augmented Generation (RAG) with Large Language Model (LLM)

Submitted By : MUHAMMED SALEEM KALLAN (22102058)

First Examiner : Ing. GINU ALUNKAL

Second Examiner : Prof. Dr. TOBIAS SCHAFFER

Introduction

- Master thesis completed at BSH Hausgeräte GmbH, Munich.
- BSH is a home appliance manufacturer.
- Developed a natural language chatbot for easier data access
- Focus on internal data platform called Stages
- Migration of data from mySQL database to Elasticsearch
- Implemented Retrieval-Augmented Generation (RAG)
- Connected the system to a Large Language Model (LLM)

Problem Statement

BSH uses the Stages platform for project and process management, which is:

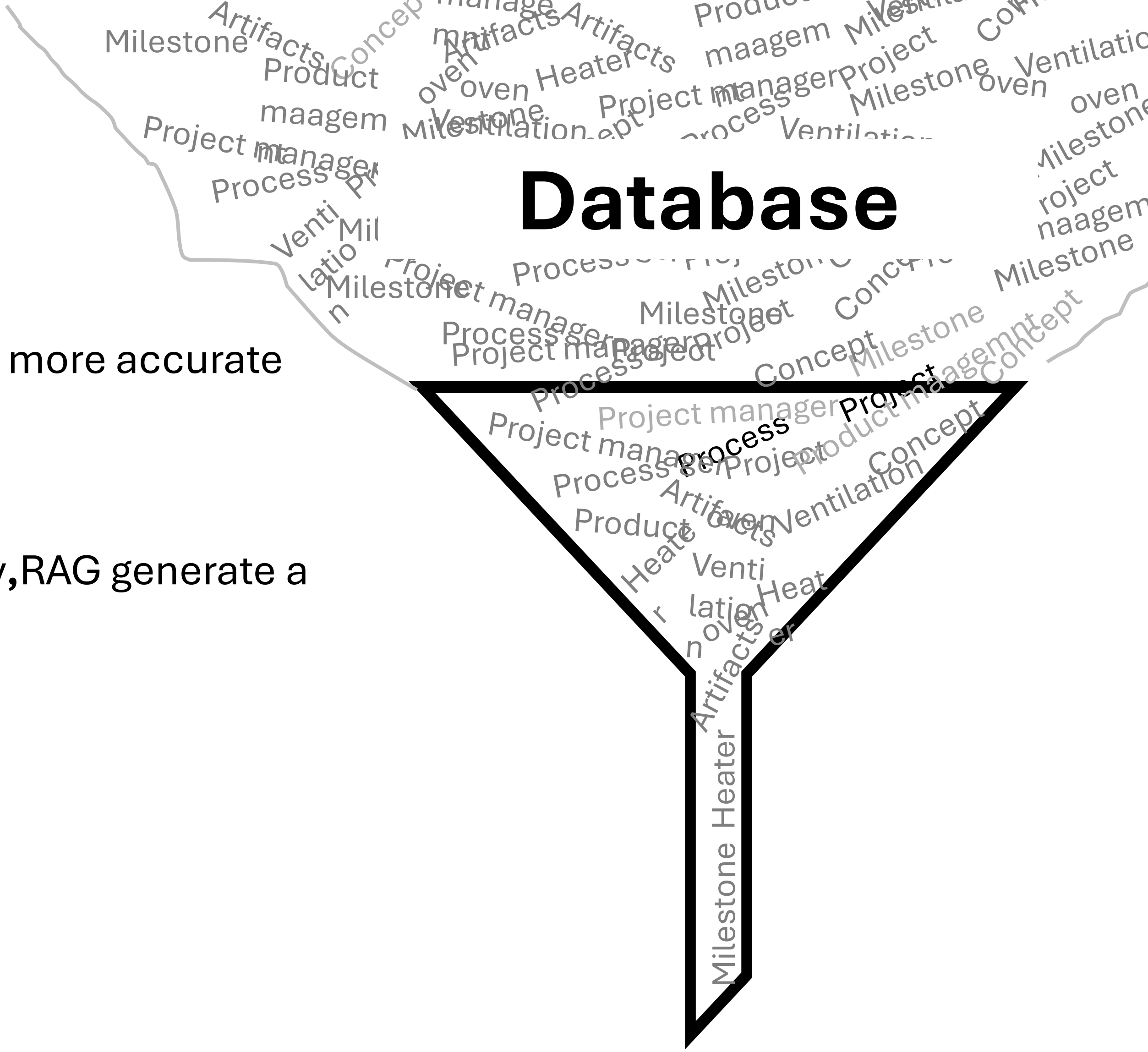
- A complex, multilayered interface
- A growing volume of technical documentation
- Time-Consuming Manual Search.
- Steep Learning Curve.

Research Objectives

- Improved Accessibility into the information
- Natural Language Understanding system
- Efficient Information Retrieval
- Enhanced Onboarding Support
- Security and Data Privacy

RAG - Retrieval Augmented Generation

- Combines search and AI text generation to give better, more accurate answers.
- Instead of the LLM "guessing" an answer from memory,RAG generate a more accurate and relevant response.



RAG - Retrieval Augmented Generation

How It Works:

1. Search:

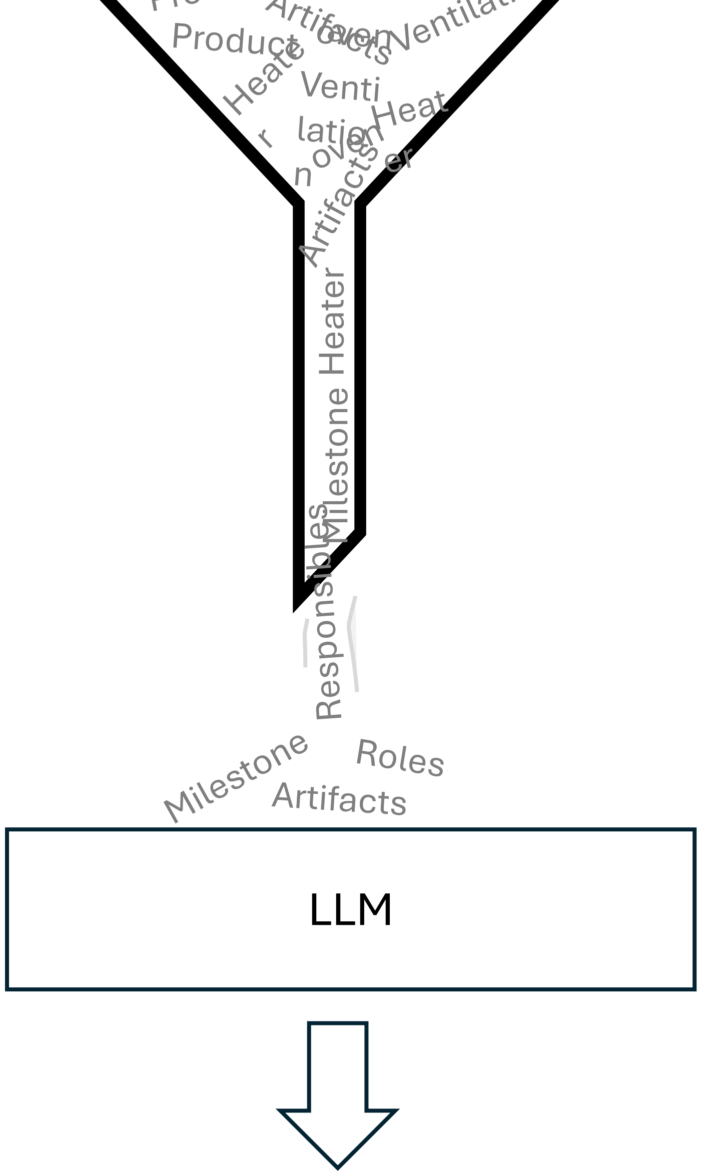
   The system search documents or text related to the question.

2. Retrieve:

   Retrieve only the relevant information using vector search, BM25 ranking, or hybrid retrieval techniques.

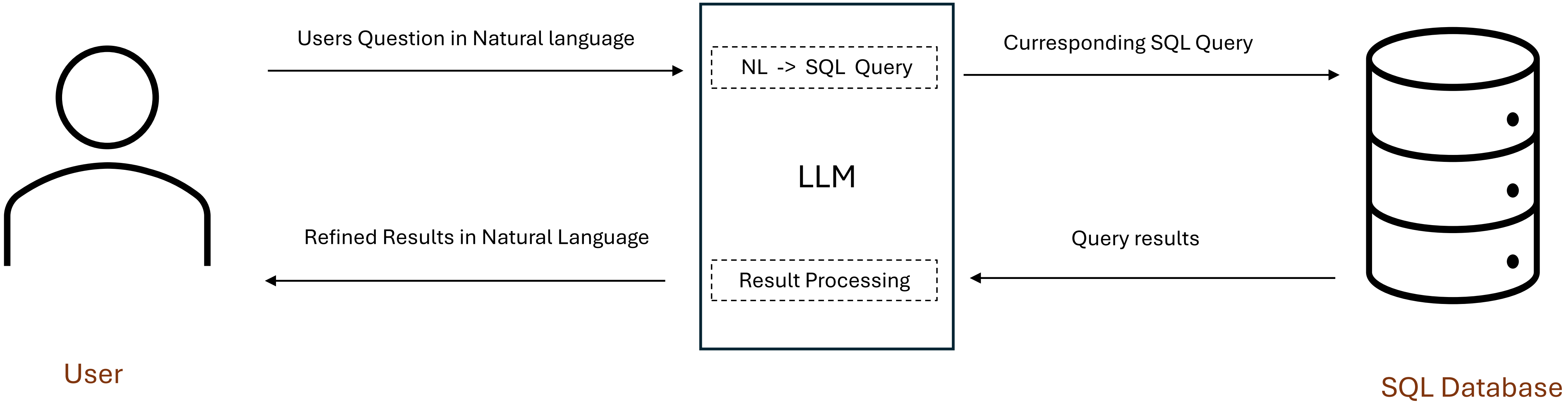
3. Generate:

   The Language model (like GPT) reads the found text and writes a response using it.



Readable Information

SQL - RAG - Chatbot



Users Question in Natural language

NL -> SQL Query

LLM

Result Processing

Refined Results in Natural Language

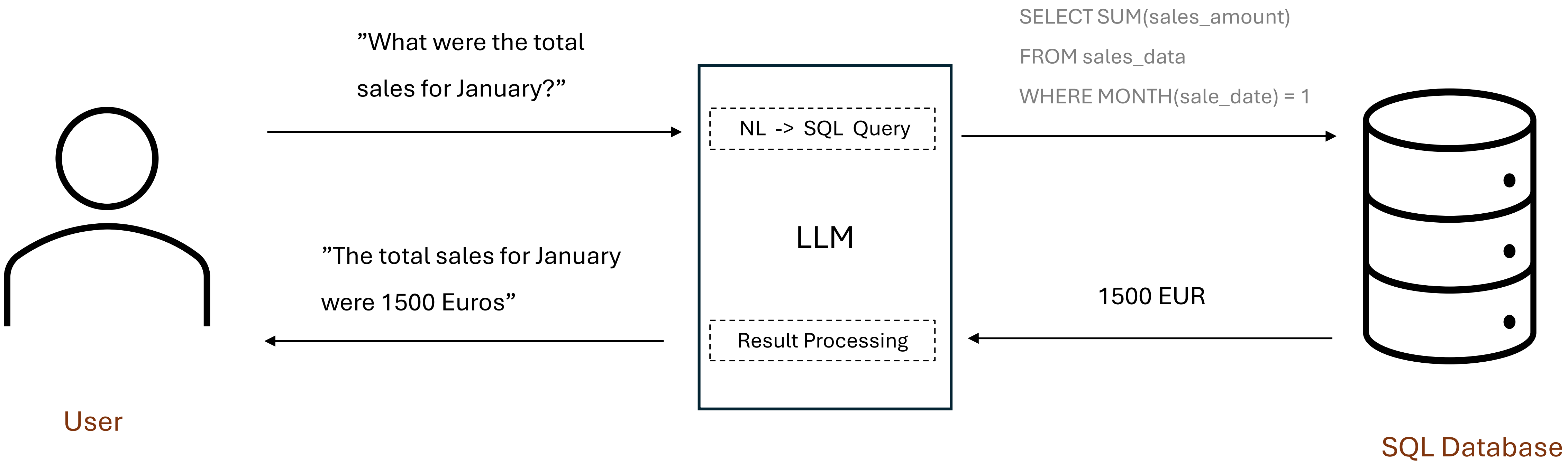
Curresponding SQL Query

Query results

User

SQL Database

SQL - RAG - Chatbot



Challenges of SQL-Based RAG Chatbots

- Lack of Semantic Understanding
- Complex Query Formulation
- Schema Dependency
- Difficulty with Unstructured Data
- Scalability and Performance Issues

Conclusion: SQL is not ideal for intelligent chatbots.

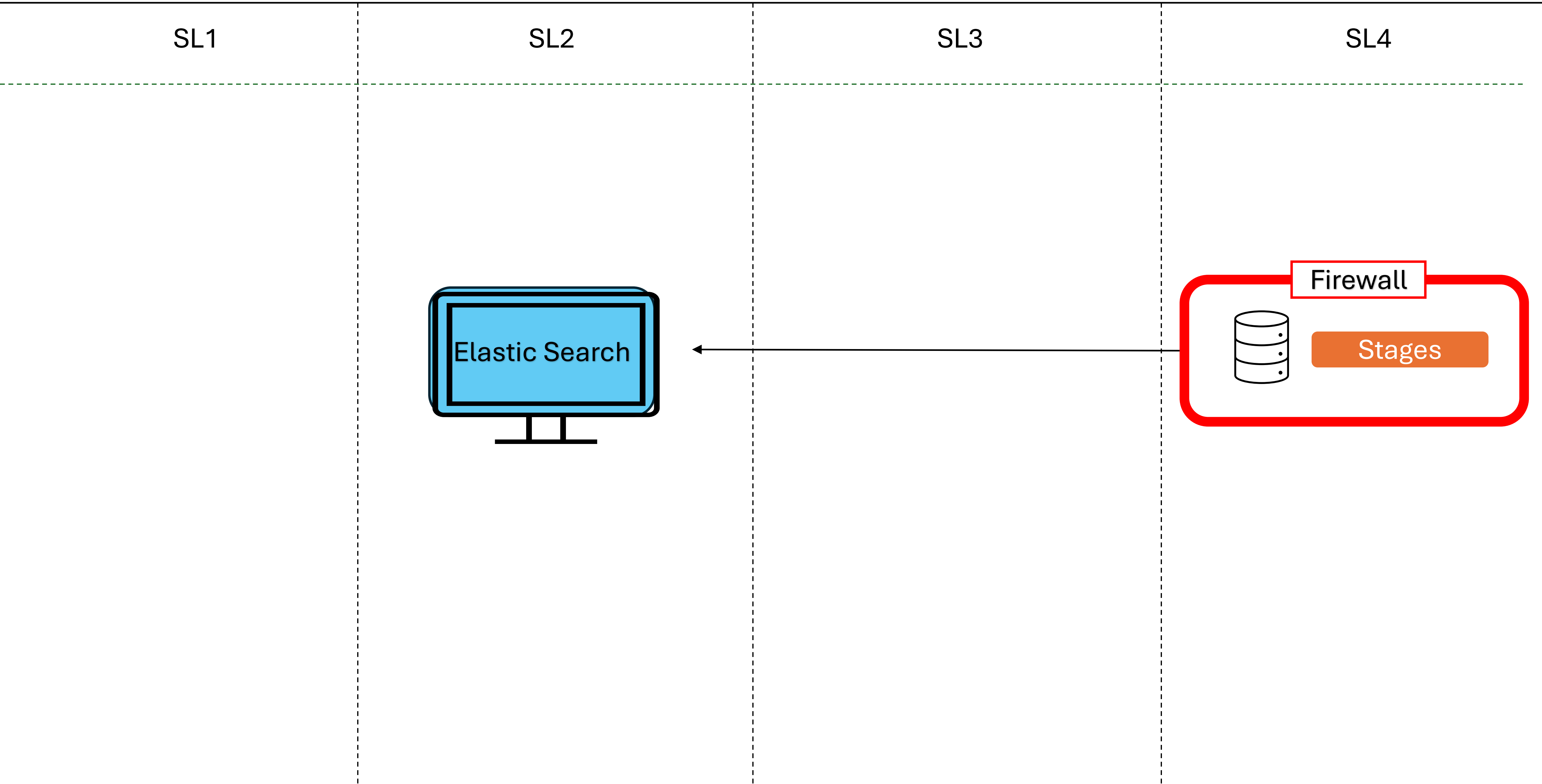
Enter Elasticsearch : designed for speed, semantic understanding, and unstructured data.

Introduction to Elasticsearch

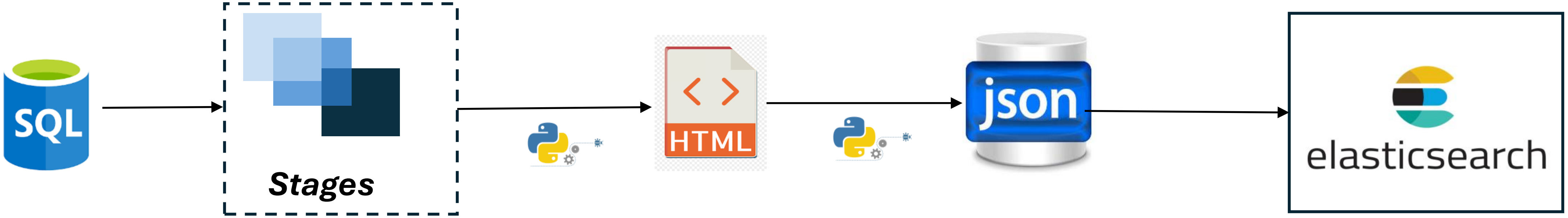
- An open-source, distributed search and analytics engine.
- Built on Apache Lucene.
- Inverted index mechanism.
- Designed for full-text search, real-time querying, and scalable data handling.

Feature	SQL Database (RDBMS)	Elasticsearch
Data Structure	Tables (rows and columns)	JSON documents
Schema	Schema based (predefined table structure)	Schema less or dynamic mapping
Storage Format	Structured data (relational)	Unstructured or semi-structured (optimized for text)
Best for	Structured data with defined relationships	Full-text search

Network Architecture



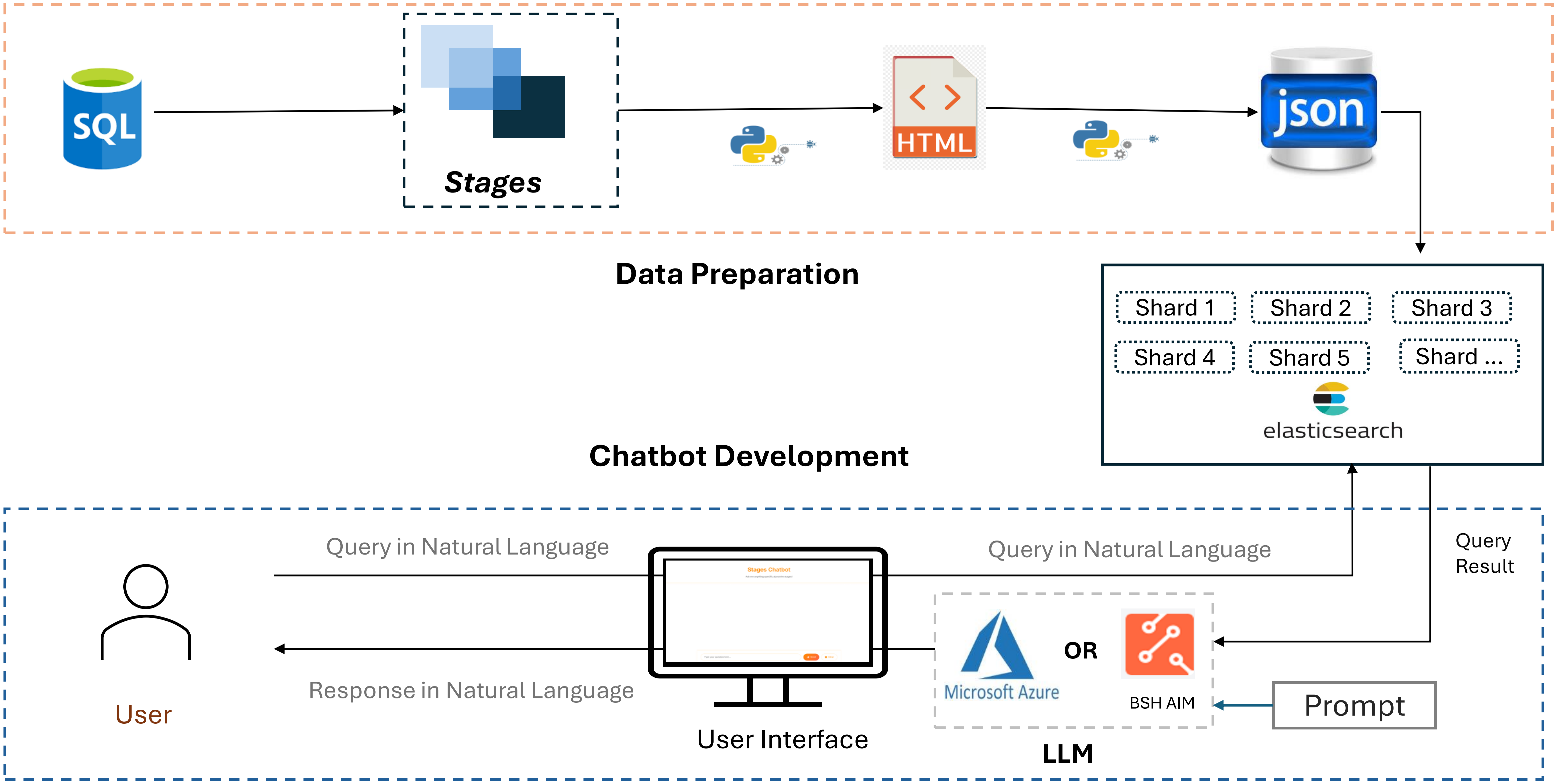
Data Preparation



- Converting HTML to JSON for Elasticsearch Ingestion
- Formatting Data for Elasticsearch
- Data Ingestion (JSON) into Elastic Search (Bulk Ingestion Using Python and REST API)
- Bulk ingestion format :

```
{ "index": { "_index": "my-index", "_id": "1" } }
{ "field1": "value1", "field2": "value2" }
{ "index": { "_index": "my-index", "_id": "2" } }
{ "field1": "value3", "field2": "value4" }
```

Architecture of the Chatbot



SQL

Stages

HTML

json

Data Preparation

Shard 1 Shard 2 Shard 3

Shard 4 Shard 5 Shard ...

elasticsearch

Chatbot Development

Query in Natural Language

Query in Natural Language

Query Result

User

Response in Natural Language

User Interface

Microsoft Azure

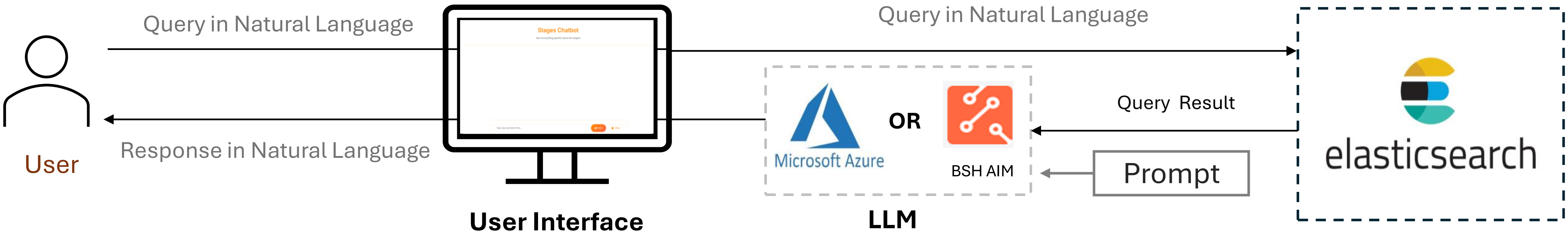
OR

BSH AIM

Prompt

LLM

Chatbot Development



User

Query in Natural Language

Response in Natural Language

Stages Chatbot

User Interface

Query in Natural Language

Microsoft Azure

OR

BSH AIM

LLM

Query Result

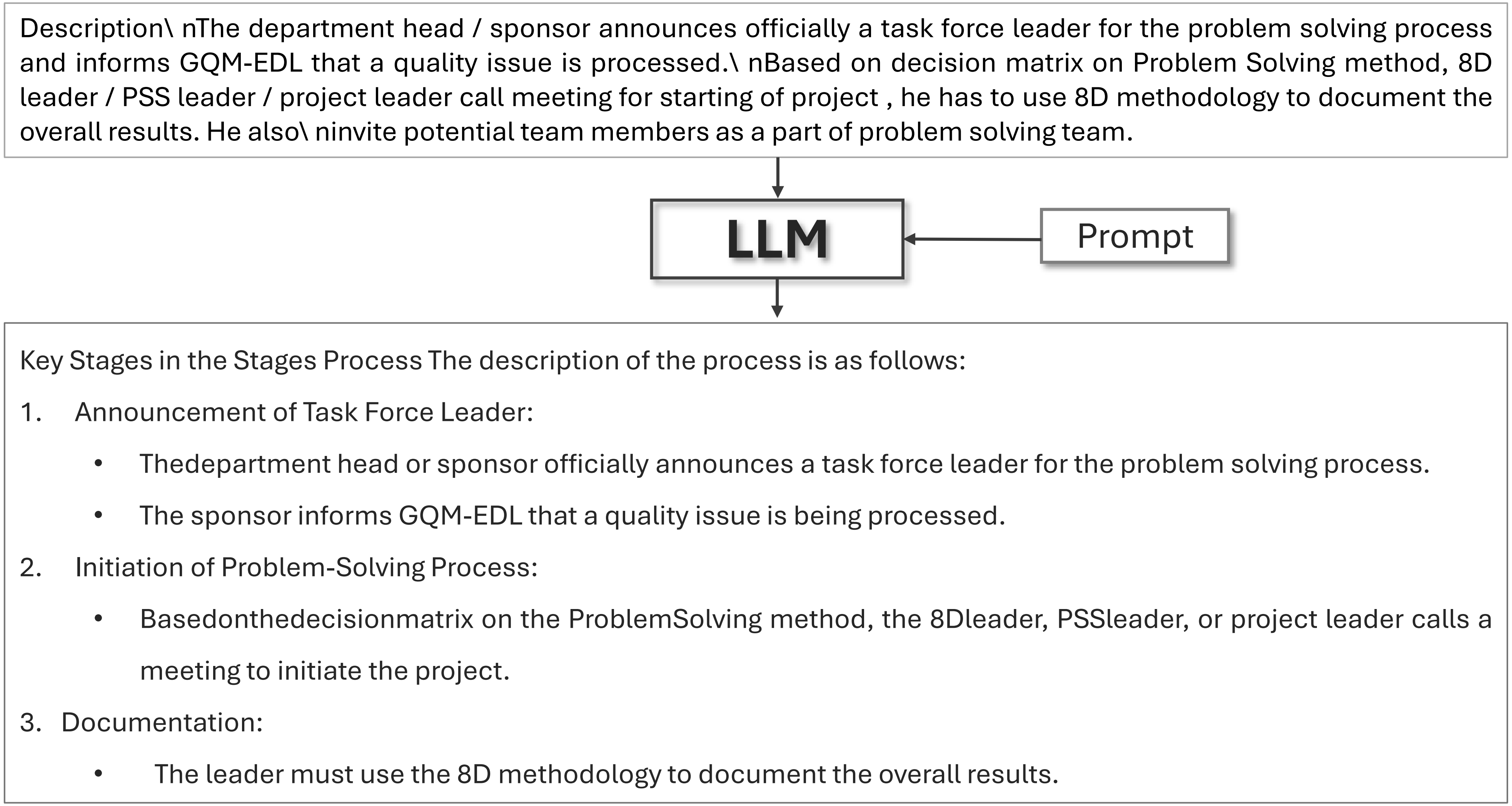
Prompt

elasticsearch

Work Flow:

1. User Query Submission
2. Query Processing & Elasticsearch Search
3. Integration with Large Language Model (LLM)
4. Response Generation & Refinement
5. Delivery to the User

Impact of Prompt on Generated responses



Description\ nThe department head / sponsor announces officially a task force leader for the problem solving process and informs GQM-EDL that a quality issue is processed.\ nBased on decision matrix on Problem Solving method, 8D leader / PSS leader / project leader call meeting for starting of project , he has to use 8D methodology to document the overall results. He also\ ninvite potential team members as a part of problem solving team.

Key Stages in the Stages Process The description of the process is as follows:

1. Announcement of Task Force Leader:
    - Thedepartment head or sponsor officially announces a task force leader for the problem solving process.
    - The sponsor informs GQM-EDL that a quality issue is being processed.
2. Initiation of Problem-Solving Process:
    - Basedonthedecisionmatrix on the ProblemSolving method, the 8Dleader, PSSleader, or project leader calls a meeting to initiate the project.
3. Documentation:
    - The leader must use the 8D methodology to document the overall results.

Testing and Results

Chatbot Testing Overview

- 42 users (Process engineers & managers at BSH).
- 350 queries collected for model evaluation.
- LLM API logs collected.

Evaluation Criteria

- BERTScore: Semantic similarity of the response.
- BLEU Score: Surface similarity of the response.

Human Evaluation

- Two independent reviewers.
- Rated on a scale of 1–5 for:
    - Relevance: Does it answer the query?
    - Coherence: Logical, consistent response?

Results

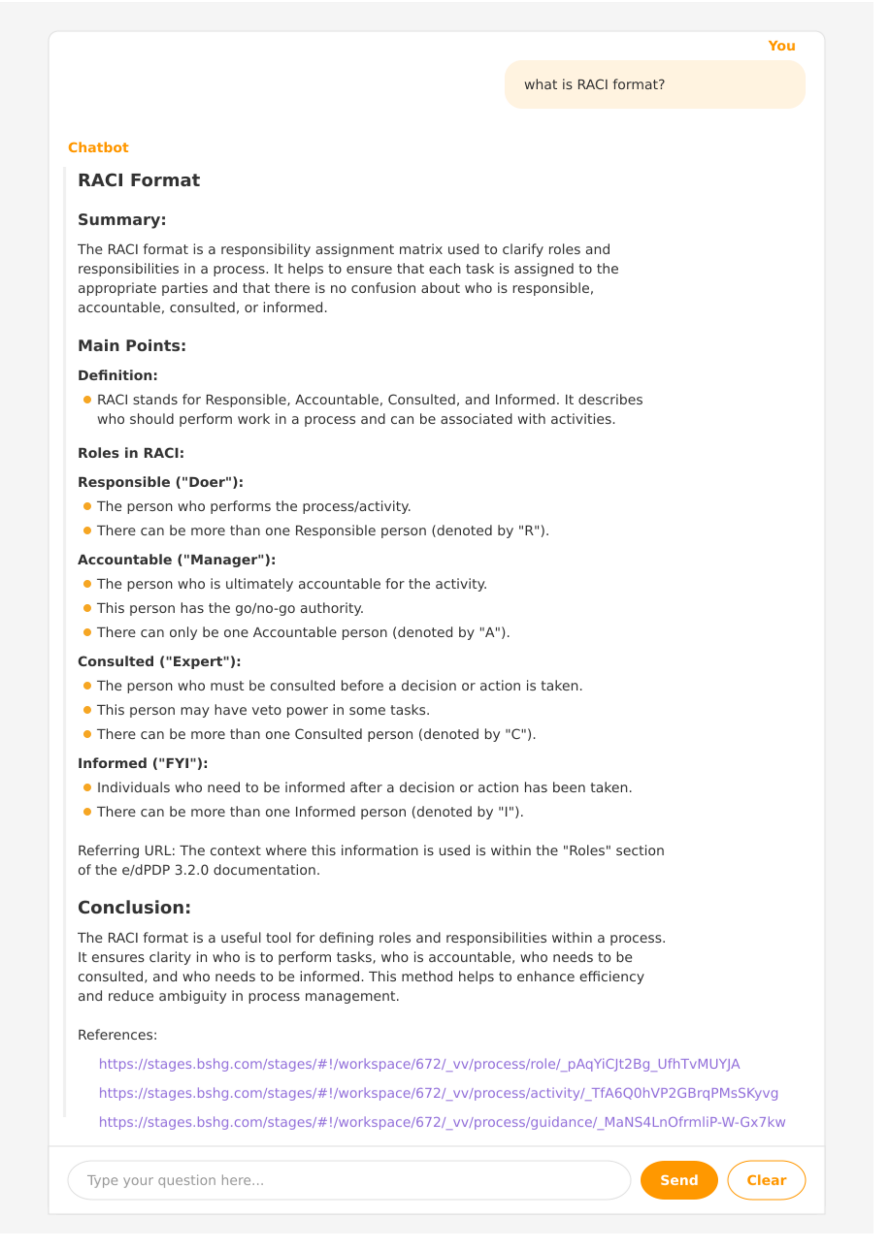
Model	BLEU Score	BERTScore	Human Rating (Relevance)	Human Rating (Coherence)	Accuracy Rate
BSH AIM GPT-3	0.61	0.84	3.5	3.6	70.5%
BSH AIM GPT-4	0.68	0.88	4.2	4.5	79.1%
Azure OpenAI GPT-3	0.58	0.82	3.4	3.5	68.0%
Azure OpenAI GPT-4	0.65	0.86	4.0	4.3	74.0%

Limitations

- Data Ingestion Approach.
- Limitations in Extracting Non-Textual Data.
- Accuracy of Retrieved Reference.

Future Potential and Discussions

- Optimizing Data Ingestion and Retrieval.
- Extracting and Incorporating Non-Textual Data.
- Strengthening Frontend and User Experience.
- Can be integrated into any website or internal system.



You

what is RACI format?

Chatbot

RACI Format

Summary:

The RACI format is a responsibility assignment matrix used to clarify roles and responsibilities in a process. It helps to ensure that each task is assigned to the appropriate parties and that there is no confusion about who is responsible, accountable, consulted, or informed.

Main Points:

Definition:

- RACI stands for Responsible, Accountable, Consulted, and Informed. It describes who should perform work in a process and can be associated with activities.

Roles in RACI:

Responsible ("Doer"):

- The person who performs the process/activity.
- There can be more than one Responsible person (denoted by "R").

Accountable ("Manager"):

- The person who is ultimately accountable for the activity.
- This person has the go/no-go authority.
- There can only be one Accountable person (denoted by "A").

Consulted ("Expert"):

- The person who must be consulted before a decision or action is taken.
- This person may have veto power in some tasks.
- There can be more than one Consulted person (denoted by "C").

Informed ("FYI"):

- Individuals who need to be informed after a decision or action has been taken.
- There can be more than one Informed person (denoted by "I").

Referring URL: The context where this information is used is within the "Roles" section of the e/dPDP 3.2.0 documentation.

Conclusion:

The RACI format is a useful tool for defining roles and responsibilities within a process. It ensures clarity in who is to perform tasks, who is accountable, who needs to be consulted, and who needs to be informed. This method helps to enhance efficiency and reduce ambiguity in process management.

References:

https://stages.bshg.com/stages/#!/workspace/672/_vv/process/role/_pAqYiCJt2Bg_UfhTvMUYJA
https://stages.bshg.com/stages/#!/workspace/672/_vv/process/activity/_TfA6Q0hVP2GBrqPMsSKyvg
https://stages.bshg.com/stages/#!/workspace/672/_vv/process/guidance/_MaNS4LnOfrmliP-W-Gx7kw

Conclusions

- Developed a chatbot with 79.1% accuracy.
- Enabled natural language based knowledge retrieval.
- Improved workflow efficiency by reducing search time.
- Ensured full compliance with company data policies.

Questions….