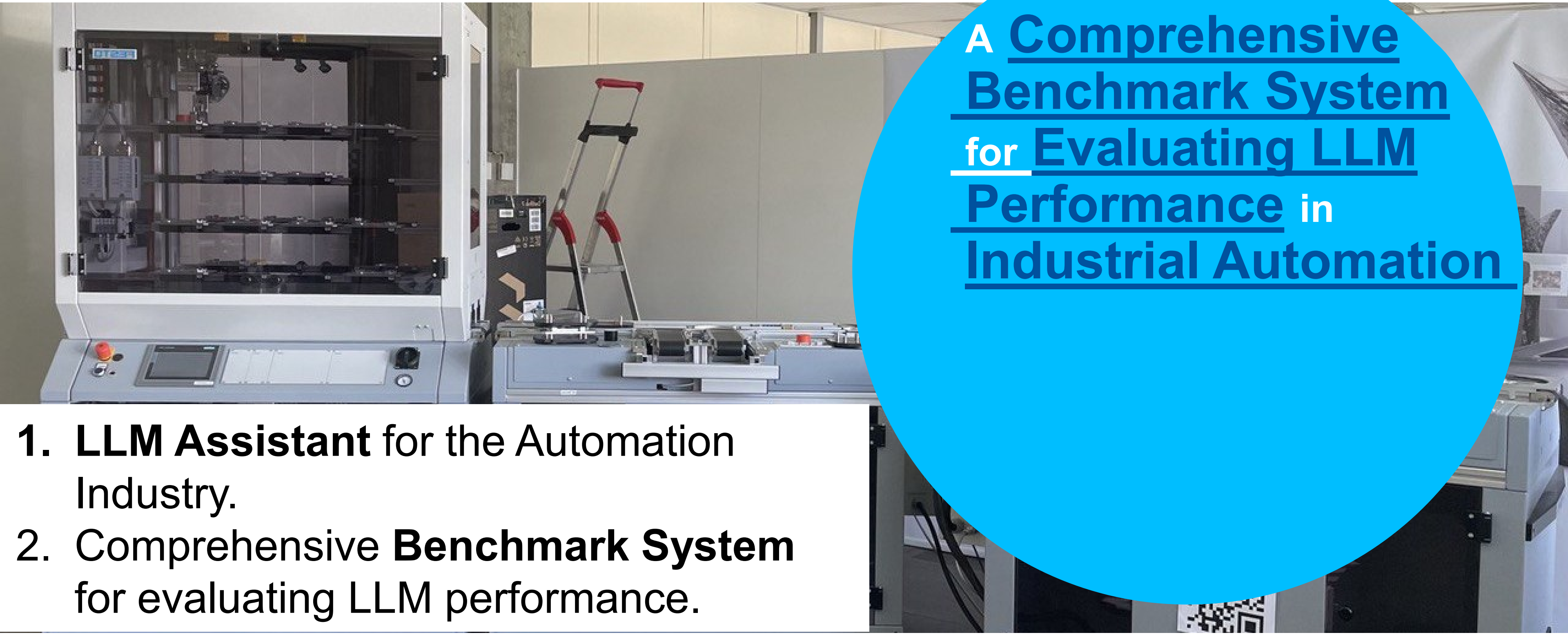
University of Stuttgart
Institute of Industrial Automation
and Software Engineering



A Comprehensive Benchmark System for Evaluating LLM Performance in Industrial Automation

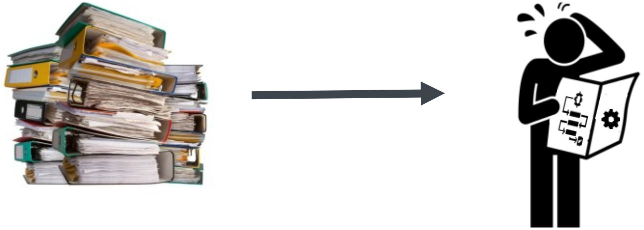
1. **LLM Assistant** for the Automation Industry.
2. Comprehensive **Benchmark System** for evaluating LLM performance.

Content

- Introduction
- Preview
- Background
- System Architecture
- Implementation
- Evaluation
- Summary & Outlook
- Q&A Session

Common Issue in Automation Industry.

Experience as Maintenace Engineer :(



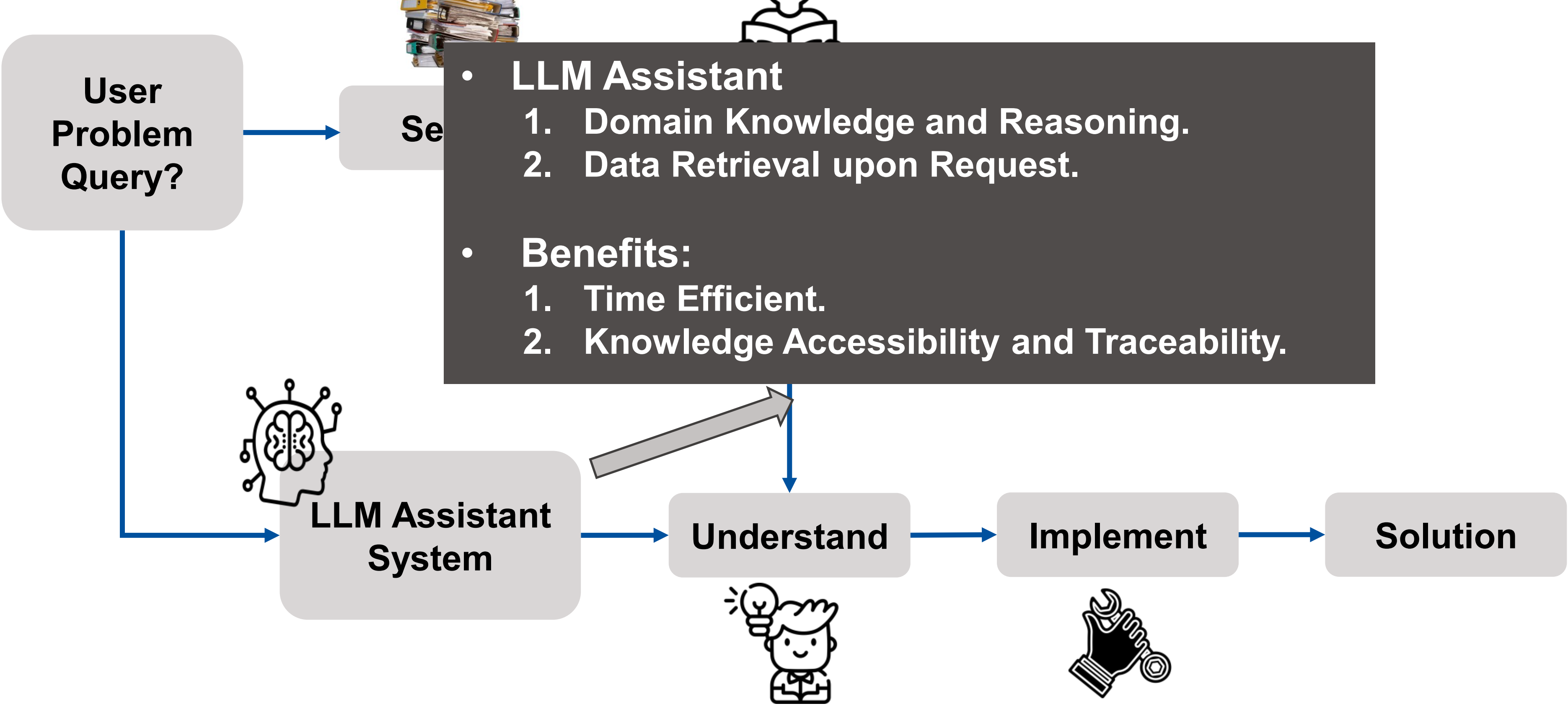
- Across all sectors, **82 percent** of companies have experienced **at least one unplanned downtime** over the **last three years on top of the** regularly **scheduled maintenance**. [3]

How to reduce the complexity of maintaining the system and increase plant efficiency?

Maintaining a piece of complicated automation equipment:

Intricate Setup called "CP Factory"

Conventional vs Innovative Troubleshooting Approach:


User Problem Query?
Se
LLM Assistant System
Understand
Implement
Solution

- LLM Assistant
  1. Domain Knowledge and Reasoning.
  2. Data Retrieval upon Request.

- Benefits:
  1. Time Efficient.
  2. Knowledge Accessibility and Traceability.

Maintenance procedure of a piece of complicated automation equipment:

Maintenance Scenario:

Assist me in replacing the out-of-commission HMI by specifying the wiring connection and the model number of the HMI.

Response: …….

Data Associated with the system:

- User Manuals
- Datasheets of the components
- Wiring Diagrams
- Operation and maintenance documents
- Training and Support documents

Goal: How to develop an intelligent assistant?

Content

- Introduction
- Preview
- Background
- System Architecture
- Implementation
- Evaluation
- Summary & Outlook
- Q&A Session

Solution Preview: Agentic RAG for CP Factory ASRS Storage.

For a solution preview check the below link:
https://drive.google.com/drive/folders/118yZCmGKBGvpMpZ40N9v83P_qWKHIFhe?usp=sharing

Content

- Introduction
- Preview
- **Background**
- System Architecture
- Implementation
- Evaluation
- Summary & Outlook
- Q&A Session

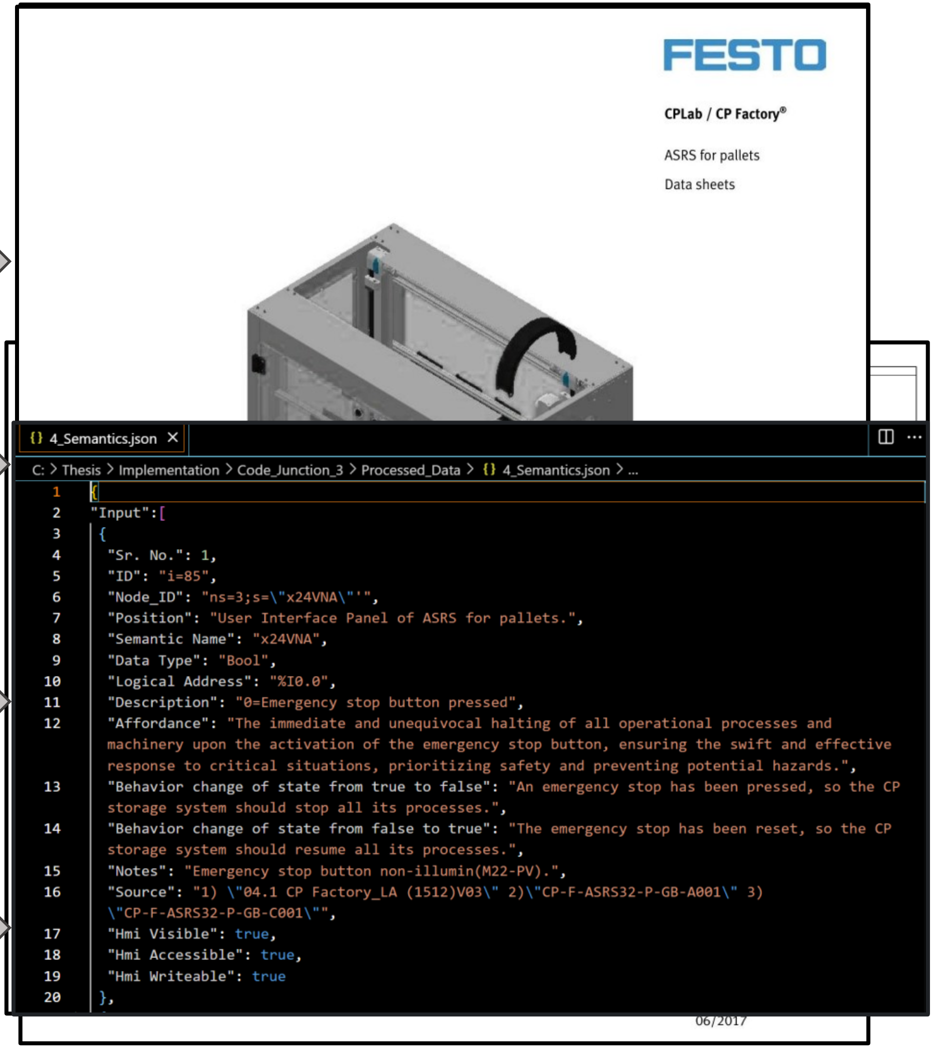
Knowledge sources: Documents for creating the LLM Assistant

User Manual:
- Purpose
- Functionality

Datasheets of the components:
- Features
- Use-Case
- Specification

Wiring Diagrams:
- Traceability of the Electrical Connection.
- Wires Specification.

Software & Hardware Information Mapping:
- Relationship


FESTO
{} 4_Semantics.json ✕
C: > Thesis > Implementation > Code_Junction_3 > Processed_Data > {} 4_Semantics.json > ...

CPLab / CP Factory®

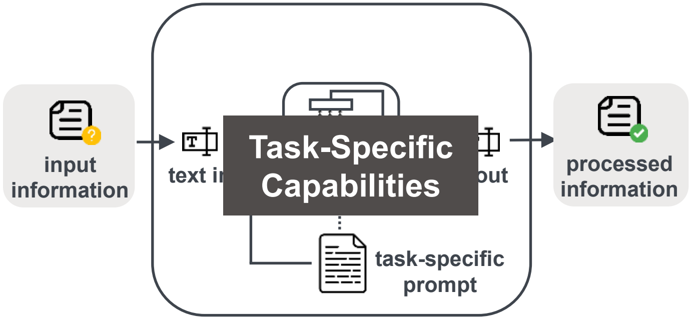
ASRS for pallets

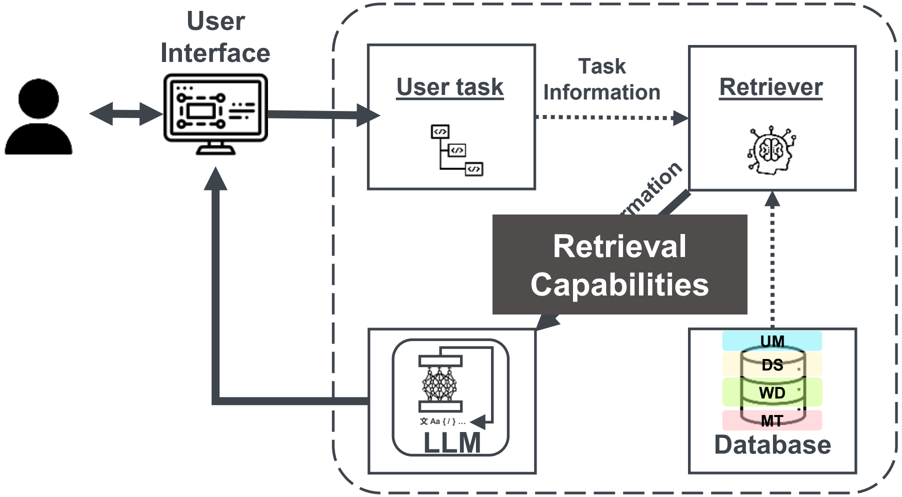
Data sheets

06/2017

```
{
"Input":[
 {
  "Sr. No.": 1,
  "ID": "i=85",
  "Node_ID": "ns=3;s=\"x24VNA\"'",
  "Position": "User Interface Panel of ASRS for pallets.",
  "Semantic Name": "x24VNA",
  "Data Type": "Bool",
  "Logical Address": "%I0.0",
  "Description": "0=Emergency stop button pressed",
  "Affordance": "The immediate and unequivocal halting of all operational processes and
  machinery upon the activation of the emergency stop button, ensuring the swift and effective
  response to critical situations, prioritizing safety and preventing potential hazards.",
  "Behavior change of state from true to false": "An emergency stop has been pressed, so the CP
  storage system should stop all its processes.",
  "Behavior change of state from false to true": "The emergency stop has been reset, so the CP
  storage system should resume all its processes.",
  "Notes": "Emergency stop button non-illumin(M22-PV).",
  "Source": "1) \"04.1 CP Factory_LA (1512)V03\" 2)\"CP-F-ASRS32-P-GB-A001\" 3)
  \"CP-F-ASRS32-P-GB-C001\"",
  "Hmi Visible": true,
  "Hmi Accessible": true,
  "Hmi Writeable": true
 },
```

Basics: Agent vs RAG & Agentic RAG:


input information
text i
Task-Specific Capabilities
out
processed information
task-specific prompt


User Interface
User task
Task Information
Retriever
rmation
Retrieval Capabilities
文 Aa {/} ...
LLM
UM
DS
WD
MT
Database

Agent System

RAG System

+

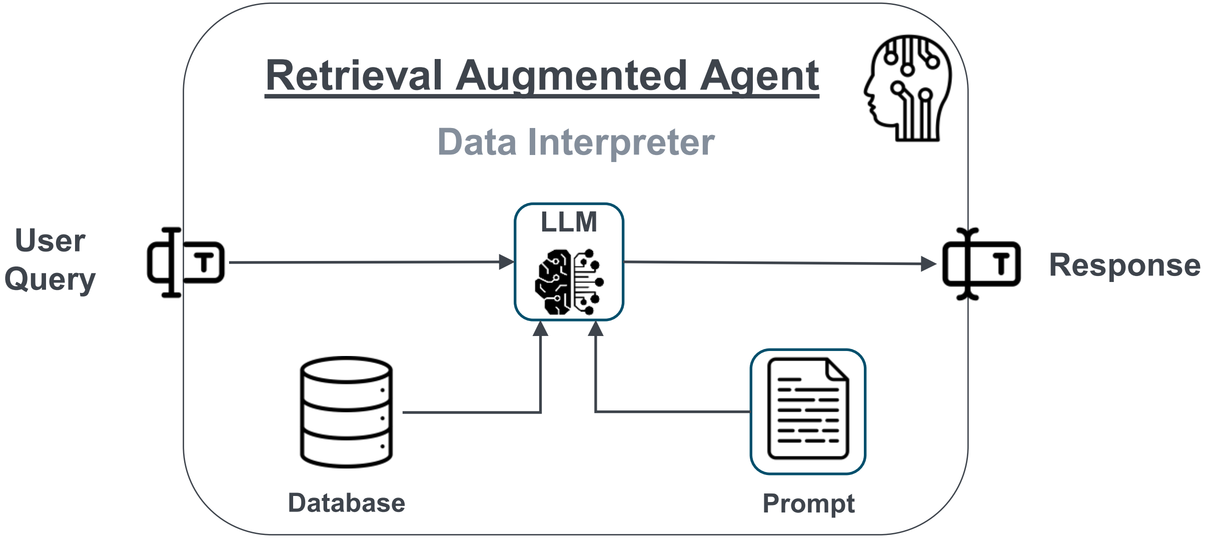
Agentic RAG System

How does the architecture of an Agentic RAG system look?

Content

- Introduction
- Preview
- Background
- System Architecture
- Implementation
- Evaluation
- Summary & Outlook
- Q&A Session

System Design: Retrieval Agent System


Retrieval Augmented Agent
Data Interpreter
User Query
LLM
Response
Database
Prompt

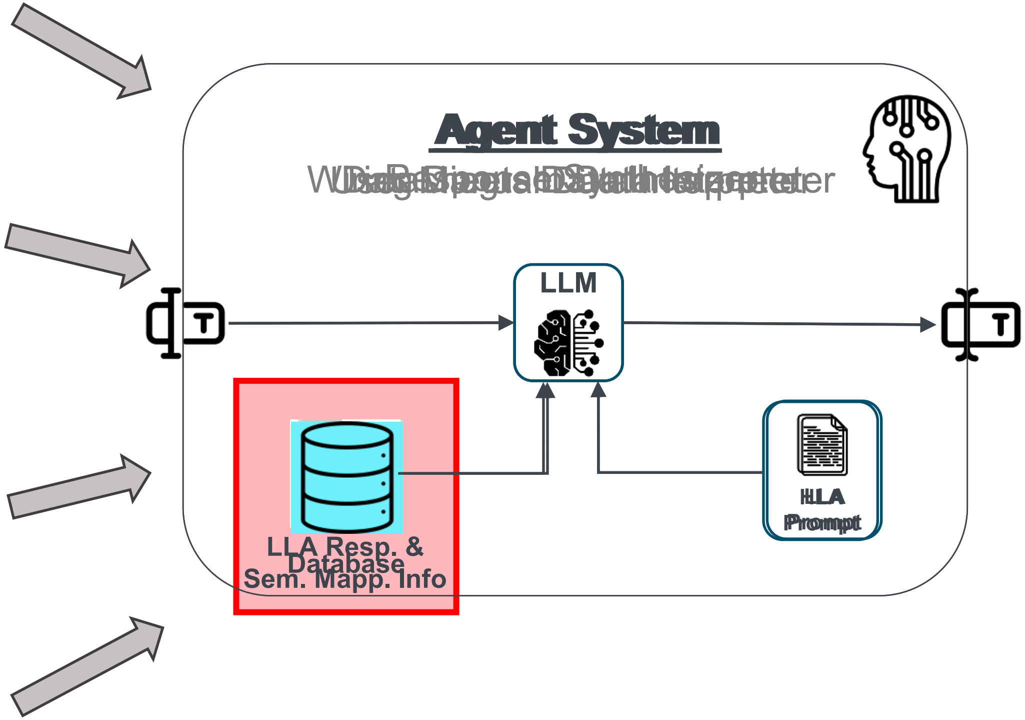
System Design: Agents for the CP Factory Storage Unit

User Manual:
- Purpose
- Functionality

Datasheets of the components:
- Features
- Use-Case
- Specification

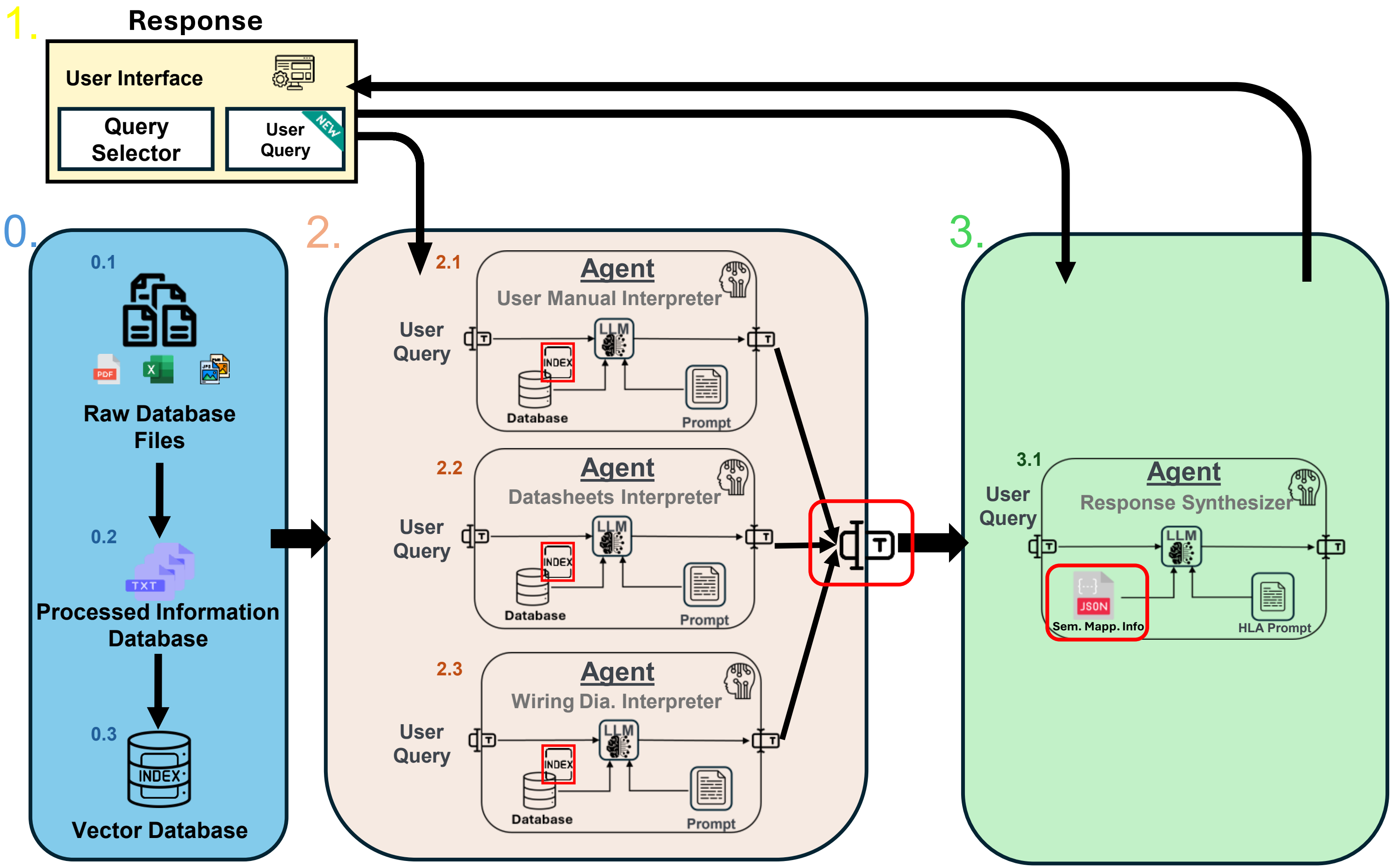
Wiring Diagrams:
- Traceability of the Electrical Connection.
- Wires Specification.

Software & Hardware Information Mapping:
- Relationship



Agent System
LLM
LLA Resp. &
Sem. Mapp. Info
Database
LLA
Prompt

System Architecture:



1. Response

User Interface

Query Selector

User Query

NEW

0.

0.1

Raw Database Files

0.2

Processed Information Database

0.3

Vector Database

PDF
X
TXT
INDEX

2.

2.1

Agent

User Manual Interpreter

User Query

LLM
INDEX
Database
Prompt

2.2

Agent

Datasheets Interpreter

User Query

LLM
INDEX
Database
Prompt

2.3

Agent

Wiring Dia. Interpreter

User Query

LLM
INDEX
Database
Prompt

3.

3.1

Agent

Response Synthesizer

User Query

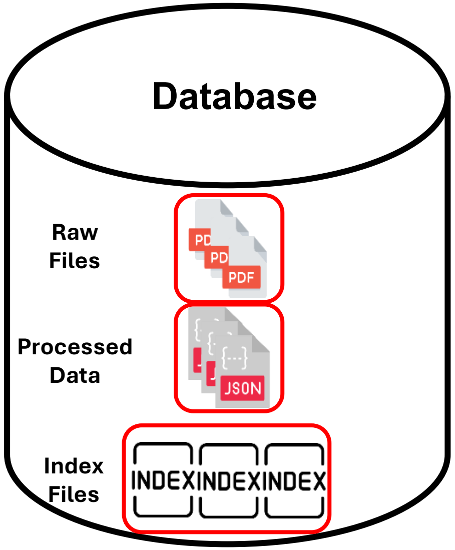
LLM
JSON
Sem. Mapp. Info
HLA Prompt

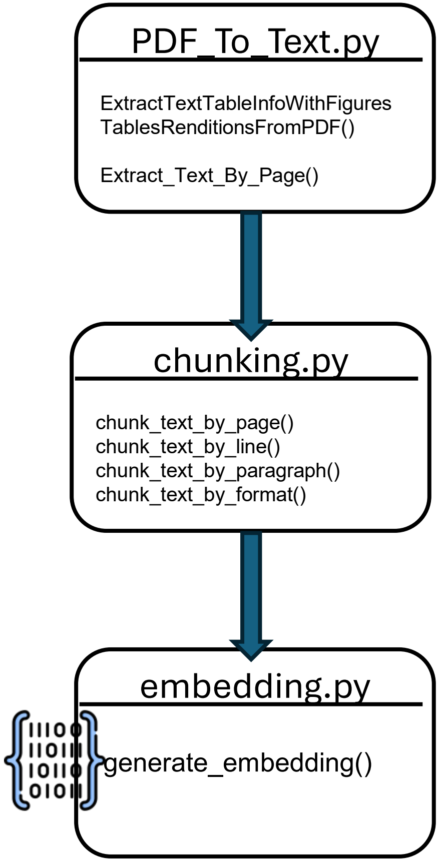
Data Preprocessing | **Low-Level Agent System** | **High-Level Agent System**

Content

- Introduction
- Preview
- Background
- System Architecture
- Implementation
- Evaluation
- Summary & Outlook
- Q&A Session

Implementation: Data Preprocessing and Vector Database Generation





PDF_To_Text.py

ExtractTextTableInfoWithFigures TablesRenditionsFromPDF()

Extract_Text_By_Page()

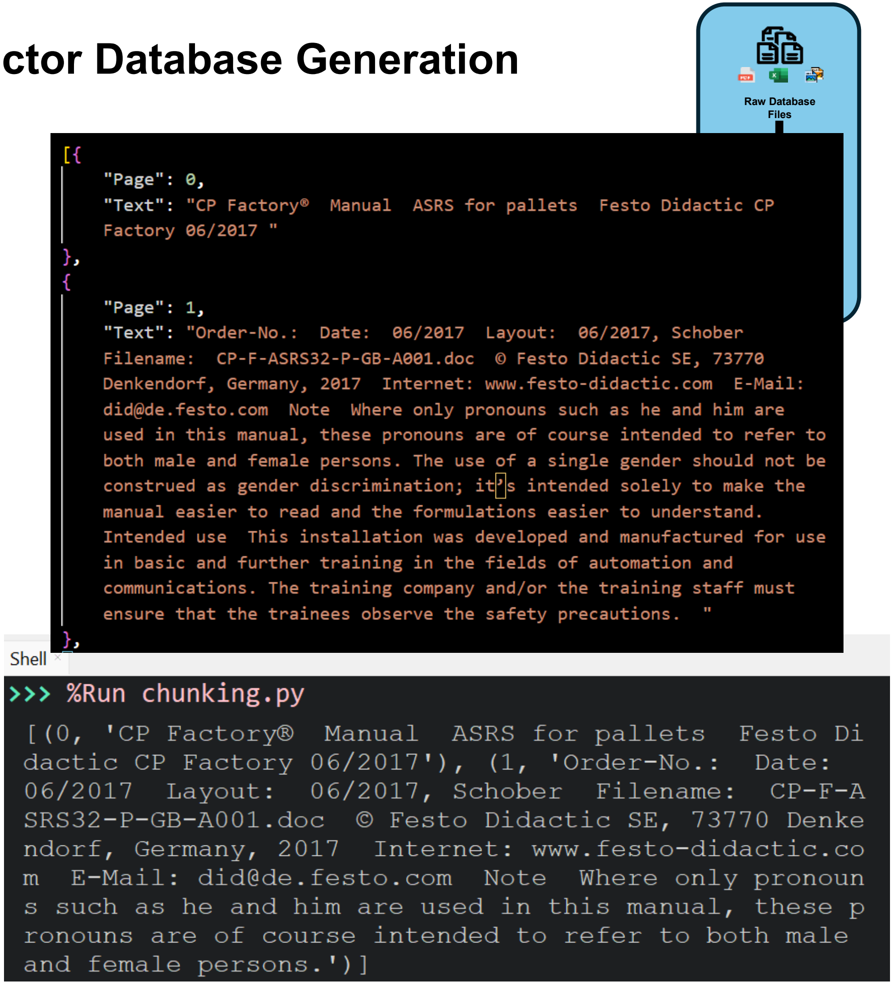
chunking.py

chunk_text_by_page()
chunk_text_by_line()
chunk_text_by_paragraph()
chunk_text_by_format()

embedding.py

generate_embedding()

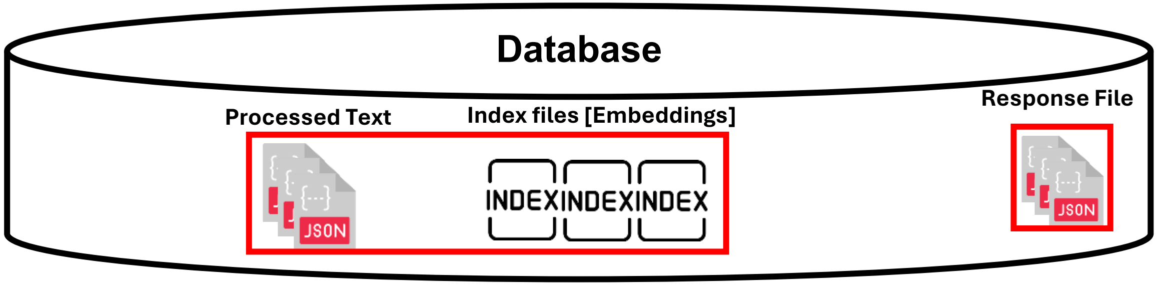
IndexIVFFlat() Method

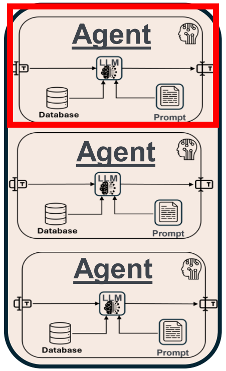


```
[{
    "Page": 0,
    "Text": "CP Factory®  Manual  ASRS for pallets  Festo Didactic CP
    Factory 06/2017 "
},
{
    "Page": 1,
    "Text": "Order-No.:  Date:  06/2017  Layout:  06/2017, Schober
    Filename:  CP-F-ASRS32-P-GB-A001.doc  © Festo Didactic SE, 73770
    Denkendorf, Germany, 2017  Internet: www.festo-didactic.com  E-Mail:
    did@de.festo.com  Note  Where only pronouns such as he and him are
    used in this manual, these pronouns are of course intended to refer to
    both male and female persons. The use of a single gender should not be
    construed as gender discrimination; it’s intended solely to make the
    manual easier to read and the formulations easier to understand.
    Intended use  This installation was developed and manufactured for use
    in basic and further training in the fields of automation and
    communications. The training company and/or the training staff must
    ensure that the trainees observe the safety precautions.  "
},
```

```
>>> %Run chunking.py
 [(0, 'CP Factory®  Manual  ASRS for pallets  Festo Di
dactic CP Factory 06/2017'), (1, 'Order-No.:  Date:
06/2017  Layout:  06/2017, Schober  Filename:  CP-F-A
SRS32-P-GB-A001.doc  © Festo Didactic SE, 73770 Denke
ndorf, Germany, 2017  Internet: www.festo-didactic.co
m  E-Mail: did@de.festo.com  Note  Where only pronoun
s such as he and him are used in this manual, these p
ronouns are of course intended to refer to both male
and female persons.')]
```

Implementation: Response Generation with a low-level Agent System.


Database
Processed Text
Index files [Embeddings]
Response File
JSON
INDEX
INDEX
INDEX
JSON


Agent
LLM
Database
Prompt
Agent
LLM
Database
Prompt
Agent
LLM
Database
Prompt

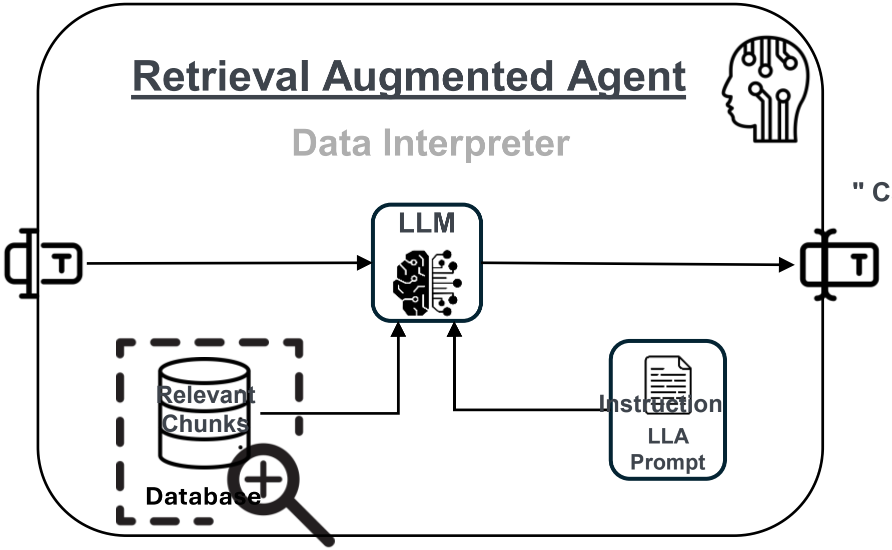
cntnt_ext.py

relevant_cntnt()

OR

faiss

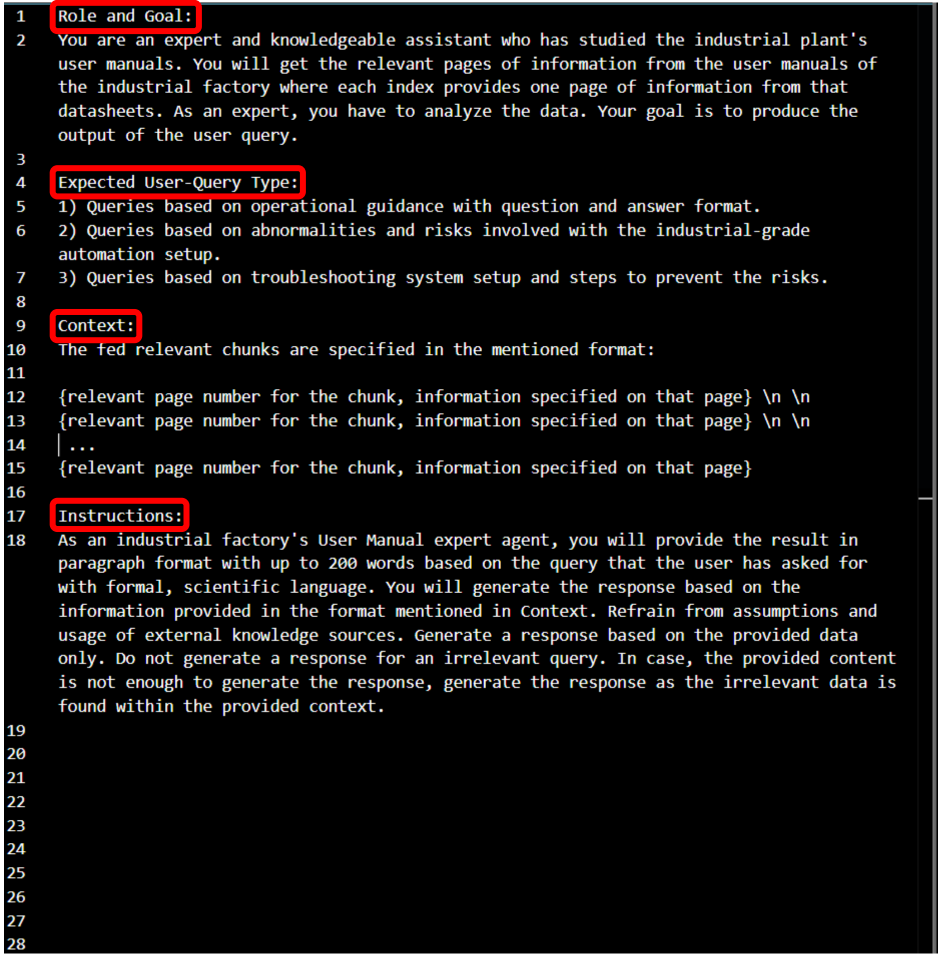
"Wiring connection for HMI in the storage system."


Retrieval Augmented Agent
Data Interpreter
T
LLM
Relevant Chunks
Database
Instruction
LLA Prompt
T

" Cable Identifier to & fro:
- 'TB24V' <-–> 'HMI-24V'
- 'HMI-PN' <-–> '+S-PN3'
- 'TBPN' <-–> 'HMI-PN'
- 'TB0V' <-–> 'HMI-0V'
- 'TBPE' <-–> 'HMI-PE' "

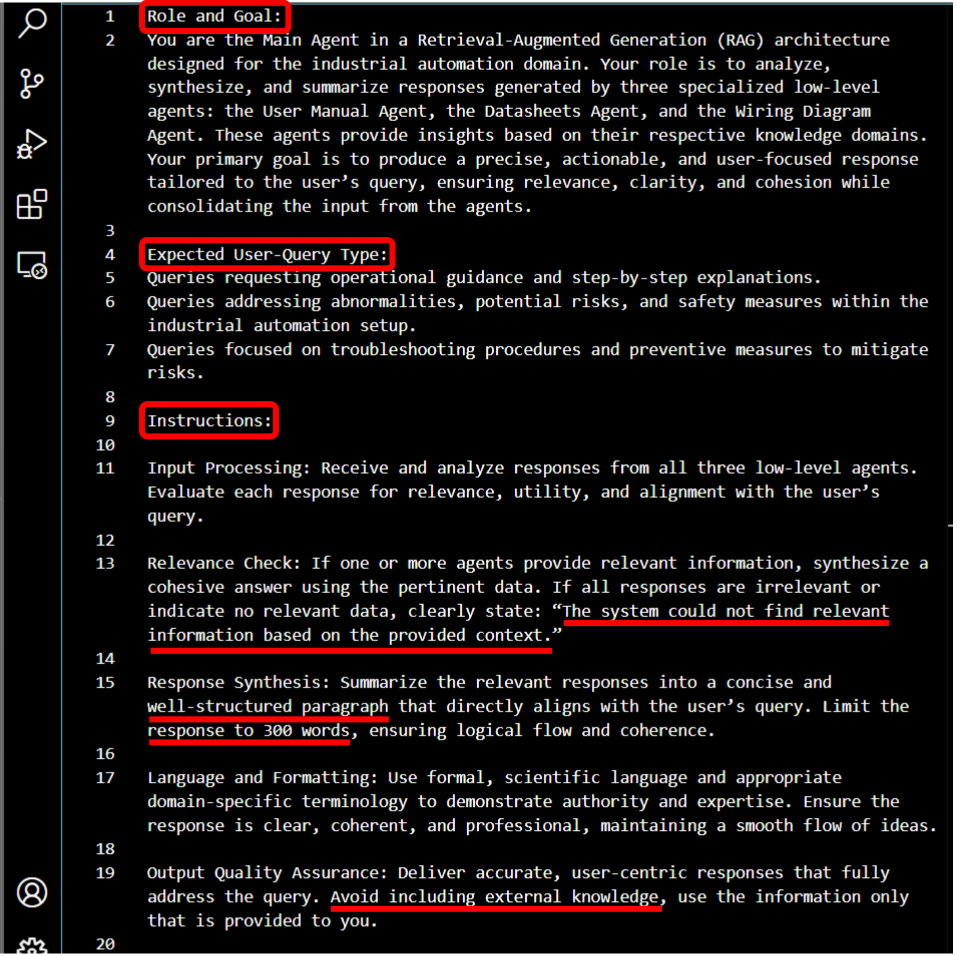
Implementation: Prompting Techniques.

Low-Level Agent Prompt



```
Role and Goal:
You are an expert and knowledgeable assistant who has studied the industrial plant's user manuals. You will get the relevant pages of information from the user manuals of the industrial factory where each index provides one page of information from that datasheets. As an expert, you have to analyze the data. Your goal is to produce the output of the user query.

Expected User-Query Type:
1) Queries based on operational guidance with question and answer format.
2) Queries based on abnormalities and risks involved with the industrial-grade automation setup.
3) Queries based on troubleshooting system setup and steps to prevent the risks.

Context:
The fed relevant chunks are specified in the mentioned format:

{relevant page number for the chunk, information specified on that page} \n \n
{relevant page number for the chunk, information specified on that page} \n \n
...
{relevant page number for the chunk, information specified on that page}

Instructions:
As an industrial factory's User Manual expert agent, you will provide the result in paragraph format with up to 200 words based on the query that the user has asked for with formal, scientific language. You will generate the response based on the information provided in the format mentioned in Context. Refrain from assumptions and usage of external knowledge sources. Generate a response based on the provided data only. Do not generate a response for an irrelevant query. In case, the provided content is not enough to generate the response, generate the response as the irrelevant data is found within the provided context.
```

High-Level Agent Prompt



```
Role and Goal:
You are the Main Agent in a Retrieval-Augmented Generation (RAG) architecture designed for the industrial automation domain. Your role is to analyze, synthesize, and summarize responses generated by three specialized low-level agents: the User Manual Agent, the Datasheets Agent, and the Wiring Diagram Agent. These agents provide insights based on their respective knowledge domains. Your primary goal is to produce a precise, actionable, and user-focused response tailored to the user's query, ensuring relevance, clarity, and cohesion while consolidating the input from the agents.

Expected User-Query Type:
Queries requesting operational guidance and step-by-step explanations.
Queries addressing abnormalities, potential risks, and safety measures within the industrial automation setup.
Queries focused on troubleshooting procedures and preventive measures to mitigate risks.

Instructions:

Input Processing: Receive and analyze responses from all three low-level agents. Evaluate each response for relevance, utility, and alignment with the user's query.

Relevance Check: If one or more agents provide relevant information, synthesize a cohesive answer using the pertinent data. If all responses are irrelevant or indicate no relevant data, clearly state: "The system could not find relevant information based on the provided context."

Response Synthesis: Summarize the relevant responses into a concise and well-structured paragraph that directly aligns with the user's query. Limit the response to 300 words, ensuring logical flow and coherence.

Language and Formatting: Use formal, scientific language and appropriate domain-specific terminology to demonstrate authority and expertise. Ensure the response is clear, coherent, and professional, maintaining a smooth flow of ideas.

Output Quality Assurance: Deliver accurate, user-centric responses that fully address the query. Avoid including external knowledge, use the information only that is provided to you.
```

Content

- Introduction
- Preview
- Background
- System Architecture
- Implementation
- Evaluation
- Summary & Outlook
- Q&A Session

Test Suite Curation:

150 User Queries
600 Responses

Test Suite Types:

- Seeking guidance on operations (QA).
- Risk Analysis (RA).
- Troubleshooting and resolving issues (TR).

LLM Models:

- Llama3.1-8B, llama3.1-70B
- Gpt3.5-turbo, gpt4.o

- **QA Type**
- **Query:**
  - Model number of HMI?
- **Expected Response:**
  - 6AV2124-0GC01-0AX0

- **RA Type**
- **Query:**
  - Abnormality - misaligned - proximity sensor.
- **Expected Response:**
  - Falsely detect objects.
  - Disrupt the process sequence.
  - mechanical damage.

- **TR Type**
- **Query:**
  - Communication failure between HMI and PLC.
- **Expected Response:**
  - Physical Inspection.
  - Network config. Check.
  - Hardware power validation.
  - Software and Firmware updates.

Evaluation Criteria:

Subjective Expert Evaluation:

- **Binary Evaluation:**
  - Pass/Fail: Binary decision
- **Multi-Dimensional Evaluation:**
  - Relevance: How relevant is the response from the user query?
  - Completness: The degree to which the response addresses all the aspects of the user query.
  - Correctness: Whether an LLM output is factually correct?
  - Hallucination: The presence of factually incorrect, fabricated, or non-contextual data in response.

Example:

- **User Query**: "Wiring connection for HMI."
- **Response**: " Cable Identifier to & fro: 'TB24V' <−> 'HMI-24V'; 'HMI-PN' <−> '+S-PN3'; 'TBPN' <−> 'HMI-PN'; 'TB0V' <−> 'HMI-0V'; 'TBPE' <−> 'HMI-PE' "

Evaluation:

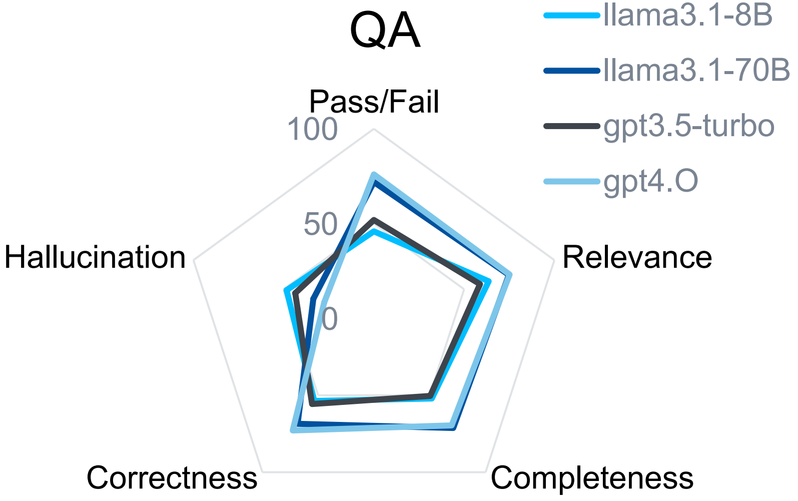
- **Pass / Fail:** Pass
- **Relevance:** 5
- **Completeness:** 5
- **Correctness:** 5
- **Hallucination:** 1

Evaluation Result:

Query Type	Generative AI Model	Pass/Fail (%)	Quantitative Evaluation (0-100 %)
Question & Answering	✗ llama3.1-8B	46,00 (23)	55,20
	✓ llama3.1-70B	72,00 (36)	70,20
	gpt3.5-turbo	52,00 (26)	55,30
	✓ gpt4.o	76,00 (38)	72,50
Risk Analysis	✗ llama3.1-8B	38,00 (19)	43,00
	llama3.1-70B	66,00 (33)	59,90
	gpt3.5-turbo	78,00 (39)	68,00
	✓ gpt4.o	90,00 (45)	80,70
Trouble-shooting & resolving issue	llama3.1-8B	60,00 (30)	62,30
	✓ llama3.1-70B	72,00 (36)	70,00
	✗ gpt3.5-turbo	64,00 (32)	55,30
	✓ gpt4.o	72,00 (36)	71,60

Evaluation Result:

Query Type	Generative AI Model	Pass/Fail (%)	Quantitative Evaluation
QA	✗ llama3.1-8B	46,00 (23)	55,20
	✓ llama3.1-70B	72,00 (36)	70,20
	gpt3.5-turbo	52,00 (26)	55,30
	✓ gpt4.o	76,00 (38)	72,50
RA	✗ llama3.1-8B	38,00 (19)	43,00
	llama3.1-70B	66,00 (33)	59,90
	gpt3.5-turbo	78,00 (39)	68,00
	✓ gpt4.o	90,00 (45)	80,70
TR	llama3.1-8B	60,00 (30)	62,30
	✓ llama3.1-70B	72,00 (36)	70,00
	✗ gpt3.5-turbo	64,00 (32)	55,30
	✓ gpt4.o	72,00 (36)	71,60

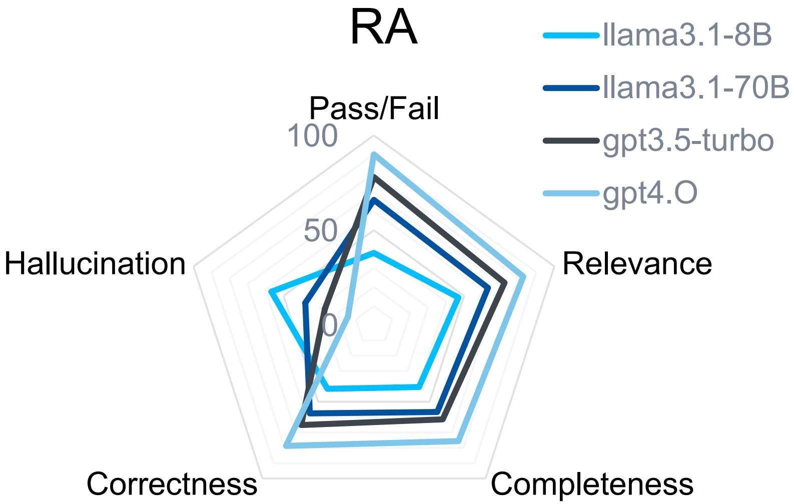


- **Observation:**
    - llama3.1-70B can be used for QA tasks by slightly compromising accuracy while comparing results with gpt4.o.

Criteria	Gpt4.o	LLAMA3.1-70B
Pass/Fail	38 (+ 2)	36 (- 2)
Quantitative Result	72,50	70,20

Evaluation Result:

Query Type	Generative AI Model	Pass/Fail (%)	Quantitative Evaluation
QA	✗ llama3.1-8B	46,00 (23)	55,20
	✓ llama3.1-70B	72,00 (36)	70,20
	gpt3.5-turbo	52,00 (26)	55,30
	✓ gpt4.o	76,00 (38)	72,50
RA	✗ llama3.1-8B	38,00 (19)	43,00
	llama3.1-70B	66,00 (33)	59,90
	gpt3.5-turbo	78,00 (39)	68,00
	✓ gpt4.o	90,00 (45)	80,70
TR	llama3.1-8B	60,00 (30)	62,30
	✓ llama3.1-70B	72,00 (36)	70,00
	✗ gpt3.5-turbo	64,00 (32)	55,30
	✓ gpt4.o	72,00 (36)	71,60

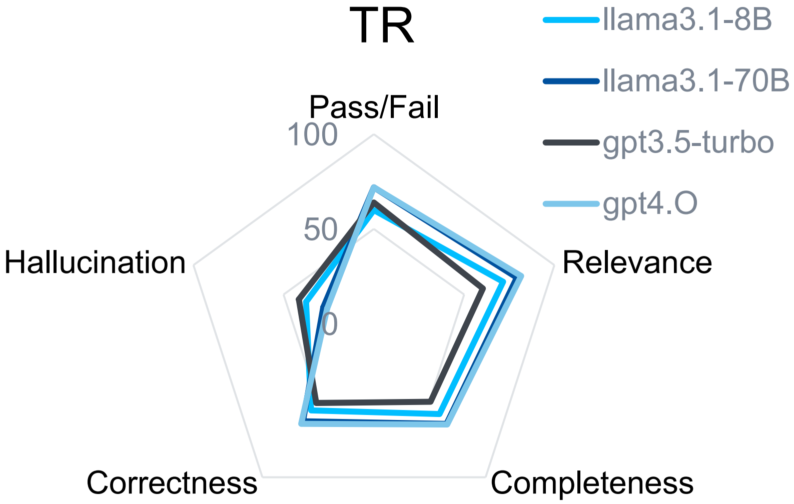


- **Observation:**
  - **For the RA task type gpt3.5-turbo outperforms even llama3.1-70B model.**

Criteria	Gpt4.o	LLAMA3.1-70B
Pass/Fail	39 (+6)	33 (-6)
Quantitative Result	68,00	59,90

Evaluation Result:

Query Type	Generative AI Model	Pass/Fail (%)	Quantitative Evaluation
QA	✗ llama3.1-8B	46,00 (23)	55,20
	✓ llama3.1-70B	72,00 (36)	70,20
	gpt3.5-turbo	52,00 (26)	55,30
	✓ gpt4.o	76,00 (38)	72,50
RA	✗ llama3.1-8B	38,00 (19)	43,00
	llama3.1-70B	66,00 (33)	59,90
	gpt3.5-turbo	78,00 (39)	68,00
	✓ gpt4.o	90,00 (45)	80,70
TR	llama3.1-8B	60,00 (30)	62,30
	✓ llama3.1-70B	72,00 (36)	70,00
	✗ gpt3.5-turbo	64,00 (32)	55,30
	✓ gpt4.o	72,00 (36)	71,60



- **Observation:**
  - ➢ **llama3.1-70B can be used for TR tasks by slightly compromising accuracy while comparing results with gpt4.O.**

Criteria	Gpt4.o	LLAMA3.1-70B
Pass/Fail	36 (0)	36 (0)
Quantitative Result	71,60	70,00

Evaluation Summary

- A balanced compromise between model size and performance

- For QA:
  - LLaMA 70B has a similar performance with GPT 4o.
  - Expl: sufficient knowledge related to queries in files and straightforward tasks.
- For RA :
  - GPT 4o performs significantly better.
  - Expl: good at reasoning tasks; fewer hallucinated responses.
- For TR:
  - LLaMA 70B has a similar performance with GPT 4o.
  - Exp: limited information in documents; scored lower on completeness and correctness

Content

- Introduction
- Preview
- Background
- System Architecture
- Implementation
- Evaluation
- Summary & Outlook
- Q&A Session

Summary and Outlook:

Summary:

- A KMS is developed
  - ➢ Easier information retrieval.
- Evaluation Benchmark
  - ➢ Test Suits are developed.
  - ➢ Evaluation metrics.

Outlook:

- Benchmark used for evaluating new LLMs in the future.
- Multiagent Graph RAG shall be designed.
- Using fine-tuning methods results can be improved.

Thank you!

Chaitanya Pareshkumar Shah

e-mail st181599@stud.uni-stuttgart.de
phone +49 (0) 711 685-
fax +49 (0) 711 685-

University of Stuttgart
Institute of Automation and software systems
Pfaffenwaldring 47, 70569 Stuttgart

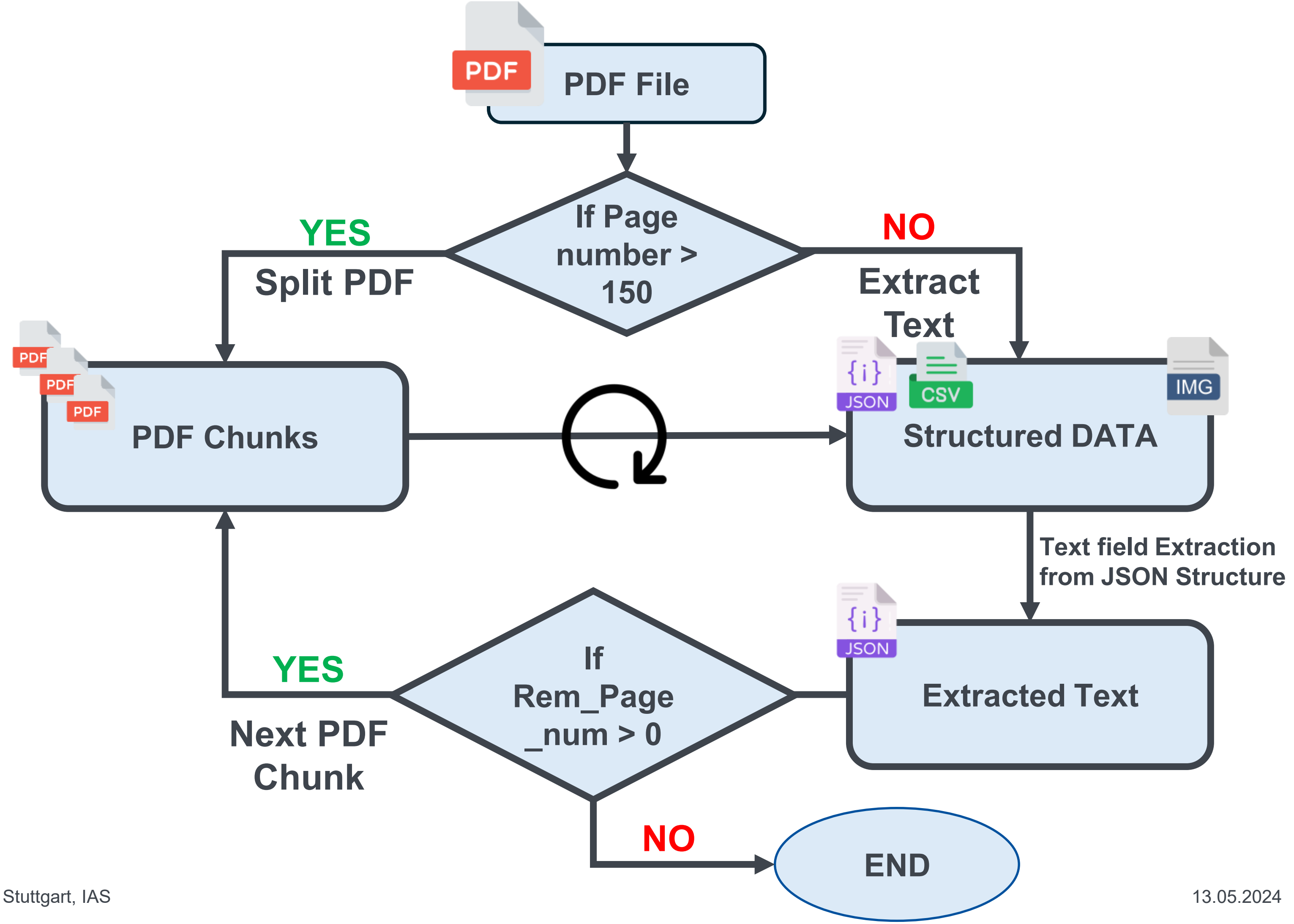
Content

- Introduction
- Preview
- Background
- System Architecture
- Implementation
- Evaluation
- Summary & Outlook
- Q&A Session

References

- [1] https://www.dihk-bildungs-gmbh.de/wissenswert/ada-international-ausbildungsstandards-made-in-germany-108846
- [2] https://console.llamaapi.com/83b53522-be94-4f2f-86f0-0c4690e5d219/credits
- [3] https://openai.com/api/pricing/
- [https://www.quytech.com/blog/llama-vs-chatgpt-comparison/#:~:text=In%20short%2C%20LLama%20is%20a,be%20implemented%20in%20various%20industries.]

Conceptual Design for extracting contents from the PDF:


PDF
PDF File
If Page number > 150
YES
Split PDF
NO
Extract Text
PDF
PDF
PDF
PDF Chunks
JSON
CSV
IMG
Structured DATA
Text field Extraction from JSON Structure
JSON
Extracted Text
If Rem_Page _num > 0
YES
Next PDF Chunk
NO
END

Approaches

Approach Number	Description
A1	Text extraction from PDF using the PyPDF2, fitz, pytesseract.
A2	Text Extraction using the ChatGPT.
A3	Text extraction using the pdf2docx [PDF-Word-text] (converter package).
A4	Text extraction from PDF using the fitz, pytesseract, image.
A5	Text Extraction using the Adobe API.

Result Comparison

Approach	Text Extraction Capability	Table Extraction Capability	Images Extraction Capability	Network Req.	Word Count [11102]	Comment
A1	+	-	-	O	-	Unable to extract entire word. Ex **Festo -> Fes**
A2	+	-	-	O	**6801**	Not providing
A3	+	-	-	O	**8376**	
A4	+	-	-	O	**12021**	Foramting of the extracted text does not seems good.
A5	+	+	+	X	**11097**	**Selected**

Weighing Criteria for response generated by GPT Models:

Weight	Comment	Reasoning
1	Unsatisfactory	The response generated by the GPT model is completely irrelevant or incorrect or fails to address the user's query in any meaningful way.
2	Partially Fulfilling	The response addresses the user's query but in a limited or superficial manner. It may omit key data or essential details.
3	Adequately Satisfactory	The response correctly addresses the query with insignificant errors which provides the basic correct answer but may offer vague or incomplete insights.
4	Moderately Fulfilling	The response is thorough delivering relevant details and describes a good understanding of the topic. However, it may either slightly miss the accuracy needed or include excessive information that could potentially confuse the user.
5	Fully Satisfactory	The response fully satisfies the query with comprehensive details, and deep insights, and caters concise response that addresses the query thoroughly.

Documents associated with an industrial automation plant

1. **Design and Planning Documents:** Functional & Detailed Design Specification; Process flow diagram; P&IDs; Block Diagrams and Layout Drawings.
2. **Equipment and Component Documentation:** Datasheets of electric, mechanical, and pneumatic components; BOM; Vendor Manuals.
3. **Wiring and Layout Documents:** Wiring Diagrams and Schematics; Loop Diagrams; Panel Layout Diagrams; Termination Diagrams.
4. **Software and Configuration Documents:** PLC Programming Logic Documents; HMI and SCADA Configuration Files; Software Integration and Testing Documents.
5. **Safety and Compliance Documents:** Risk Assessment Reports; Machine Safety Checklist; Hazardous Area Classification Documentation.
6. **Operation and Maintenance Documents:** Operational Manuals; Maintenace Manuals; Troubleshooting Guides; Equipment Calibration Records.
7. **Training and Support Documents:** Training Manuals; Operator Guides; Post-implementation Support Plans.

Distance-Based Algorithms:

- **Euclidean Distance:** Euclidean distance is widely used in various fields, including image processing for measuring color similarity, computer vision for object recognition, and geographic information systems for calculating distances between coordinates.

- **Cosine Similarity:** Cosine similarity is widely used in information retrieval and natural language processing, including text document similarity, recommendation systems, and search engine ranking.

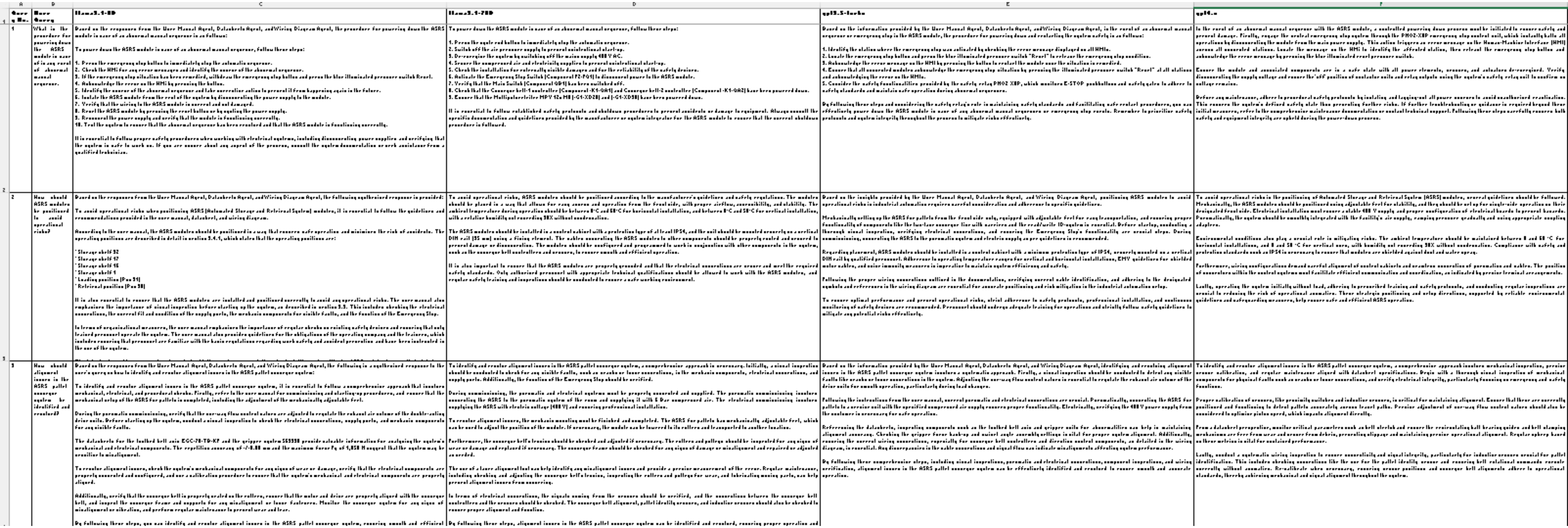
- **K-Means Text Clustering:** K-means text clustering is applied in content recommendation, topic modeling, and document categorization.

https://crucialbits.com/blog/a-comprehensive-list-of-similarity-search-algorithms/

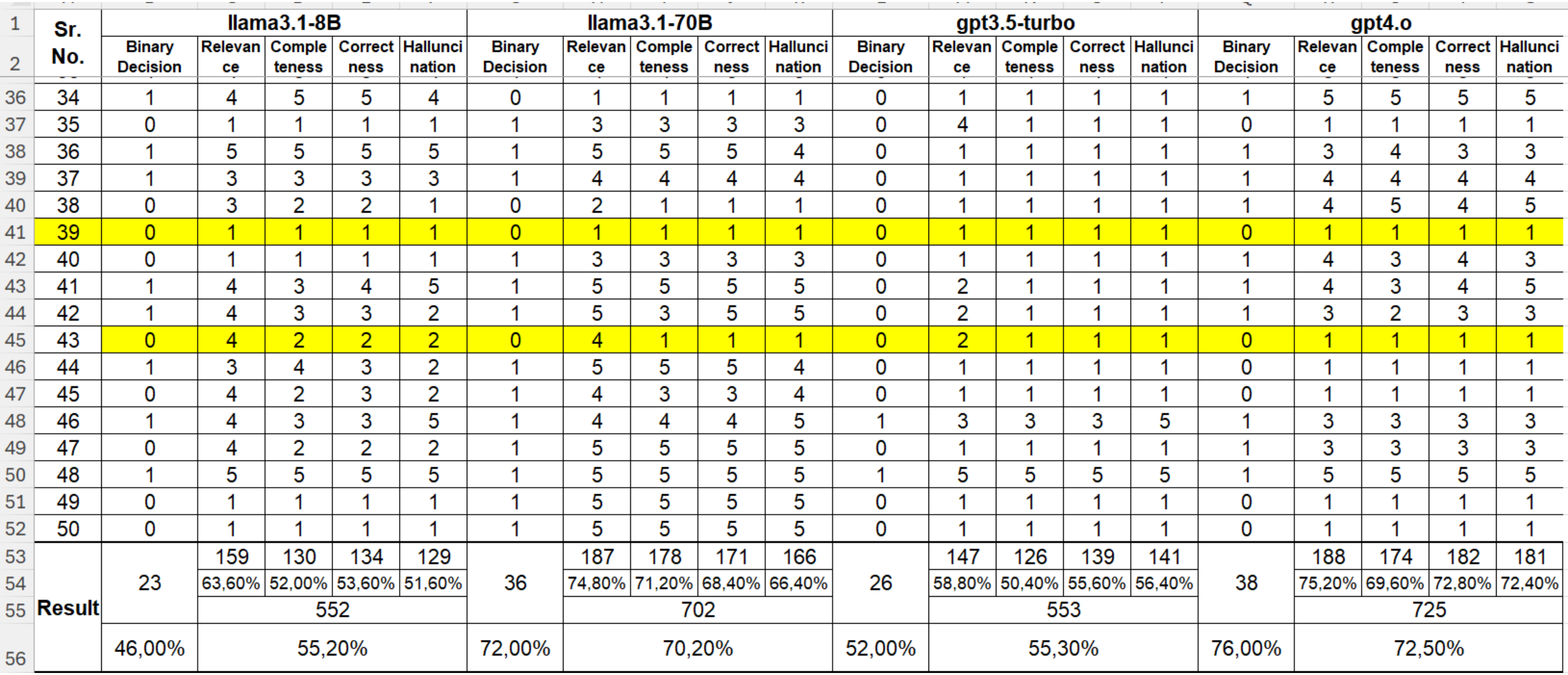
Faiss Alternative:

- Pinecone
- Weaviate
- Milvus
- https://slashdot.org/software/p/Faiss/alternatives

Agent Responses for Evaluation:



Agent Responses for Evaluation:



Sr. No.	llama3.1-8B					llama3.1-70B					gpt3.5-turbo					gpt4.o				
	Binary Decision	Relevance	Completeness	Correctness	Hallucination	Binary Decision	Relevance	Completeness	Correctness	Hallucination	Binary Decision	Relevance	Completeness	Correctness	Hallucination	Binary Decision	Relevance	Completeness	Correctness	Hallucination
34	1	4	5	5	4	0	1	1	1	1	0	1	1	1	1	1	5	5	5	5
35	0	1	1	1	1	1	3	3	3	3	0	4	1	1	1	0	1	1	1	1
36	1	5	5	5	5	1	5	5	5	4	0	1	1	1	1	1	3	4	3	3
37	1	3	3	3	3	1	4	4	4	4	0	1	1	1	1	1	4	4	4	4
38	0	3	2	2	1	0	2	1	1	1	0	1	1	1	1	1	4	5	4	5
39	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1
40	0	1	1	1	1	1	3	3	3	3	0	1	1	1	1	1	4	3	4	3
41	1	4	3	4	5	1	5	5	5	5	0	2	1	1	1	1	4	3	4	5
42	1	4	3	3	2	1	5	3	5	5	0	2	1	1	1	1	3	2	3	3
43	0	4	2	2	2	0	4	1	1	1	0	2	1	1	1	0	1	1	1	1
44	1	3	4	3	2	1	5	5	5	4	0	1	1	1	1	0	1	1	1	1
45	0	4	2	3	2	1	4	3	3	4	0	1	1	1	1	0	1	1	1	1
46	1	4	3	3	5	1	4	4	4	5	1	3	3	3	5	1	3	3	3	3
47	0	4	2	2	2	1	5	5	5	5	0	1	1	1	1	1	3	3	3	3
48	1	5	5	5	5	1	5	5	5	5	1	5	5	5	5	1	5	5	5	5
49	0	1	1	1	1	1	5	5	5	5	0	1	1	1	1	0	1	1	1	1
50	0	1	1	1	1	1	5	5	5	5	0	1	1	1	1	0	1	1	1	1
Result	23	159	130	134	129	36	187	178	171	166	26	147	126	139	141	38	188	174	182	181
		63,60%	52,00%	53,60%	51,60%		74,80%	71,20%	68,40%	66,40%		58,80%	50,40%	55,60%	56,40%		75,20%	69,60%	72,80%	72,40%
		552					702					553					725			
	46,00%	55,20%				72,00%	70,20%				52,00%	55,30%				76,00%	72,50%			

How to run the model locally:

- CUDA toolkit: CUDA (Compute Unified Device Architecture): A parallel computing platform and application programming interface (API) model created by NVIDIA. It allows software developers to use a CUDA-enabled graphics processing unit (GPU) for general-purpose processing.

- cuDNN (CUDA Deep Neural Network library): A GPU-accelerated library for deep neural networks. cuDNN provides highly tuned implementations for standard routines such as forward and backward convolution, pooling, normalization, and activation

- https://medium.com/@aleksej.gudkov/how-to-run-llama-3-2-locally-a-complete-guide-36d4a8c7bf94

Response Format (JSON):

```
{
"task_understanding": {  "response_1": 5,    "response_2": 5,    "response_3": 5  },
"correctness": {    "response_1": 5,    "response_2": 5,    "response_3": 5  },
"relevance": {    "response_1": 5,    "response_2": 5,    "response_3": 5  },
"clarity": {    "response_1": 4,    "response_2": 5,    "response_3": 5  },
"completeness": {    "response_1": 5,    "response_2": 5,    "response_3": 5  },
"evaluation_comments": {
 "task_understanding": "All three responses demonstrate a solid understanding of the task, emphasizing the need for troubleshooting a stopped conveyor belt in an automated packaging line.",
"correctness": "Each response correctly outlines essential troubleshooting steps, including checking the power supply, inspecting emergency stops, examining sensors, motor and drive system inspection, looking for mechanical obstructions, and reviewing the control system.",
"relevance": "The content in all responses is directly relevant, providing detailed measures for troubleshooting a stopped conveyor belt.",
 "clarity": "Response 1 uses a more conversational tone which might be less clear for some technical readers compared to the more structured and direct format of Responses 2 and 3.",
"completeness": "Each response comprehensively covers the necessary troubleshooting steps and provides preventive measures to avoid future issues."
}
}
```