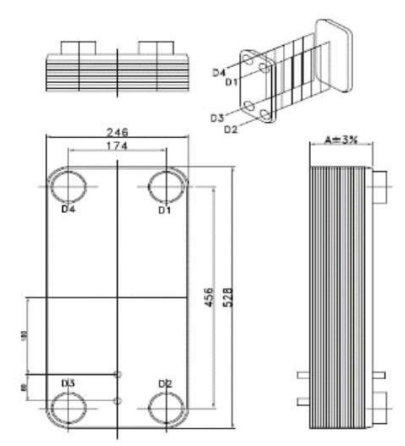


AI Information Extraction – Ask a Data Sheet

Christian Möller , Project Researcher, Novia UAS
Lamin Jatta , Project Researcher, Novia UAS
Christoffer Björkskog, Project Researcher, Novia UAS

Virtual Sea Trial

Demo – Ask a Data Sheet

Heat-Exchanger Specification						
RISK FACTOR	CUSTOMER NAME		HE Company	REF.	ITEM NO.	REMARKS
	NOVIA RDI					
	MODEL					
4	HE-SW					
DATE, PLACE						
Turku FIN						
The HE Company						
ALL DIMENSIONS ARE IN MILLIMETERS						
						
SIDE	MEDIA	SP. HEAT CAPACITY	INLET TEMP.	OUTLET TEMP.	FLOW RATE.	LIQUID VOL.
1	Water	4190 J/kg/K	68 °C	89 °C	46 kg/s	400 dm3
2	Water	4190 J/kg/K	91 °C	74 °C	56 kg/s	400 dm3

This demo showcases how to extract information from a pdf and make it searchable. Data sheets containing technical specification of components, like a heat exchanger, are the focus of this demo. To provide factual information we develop a **Retrieval Augmented Generation (RAG)** system that is based on the llama index JSONQueryEngine.

The **two-step process** start with extracting relevant information from a pdf and use this context information to augment the user query. The LLM is able to use the context and the user query to reliably provide correct answers to natural language queries.

Highlights

- System provides **bounding boxes** for the retrieved information enabling a user-centered design
- Dynamic **horizontal alignment** of the extracted information based on the query



Virtual Sea Trial

LlamaIndex JSONQueryEngine

JSONQueryEngine uses information structured as a JSON to perform **RAG**. At the core of the engine lies the JSON _value and the JSON_schema that needs to be provided.

- **JSON_value**: represents the information that is to be queried and constitutes the factual context of the user query. Needs to match the schema!
- **JSON_schema**: schema of the JSON_value. Should contain descriptions for each property.

Given a query, a LLM generates a JSONPath based on the provided JSON_schema (-> a proper description of the properties in the schema is important). The JSONPath that is generated can be used to find the relevant information for the query in the JSON_value. An LLM can be used to create a natural language response from the relevant information in the JSON_value.

JSONPath defines syntax to query JSONs

Llama JSONQueryEngine can be run in chat mode. This way it 'remembers' the chat history. False per default.

JSONQueryEngine works well if:

- Information can be represented as JSON
- JSON schema can be defined
- There are no relations between properties



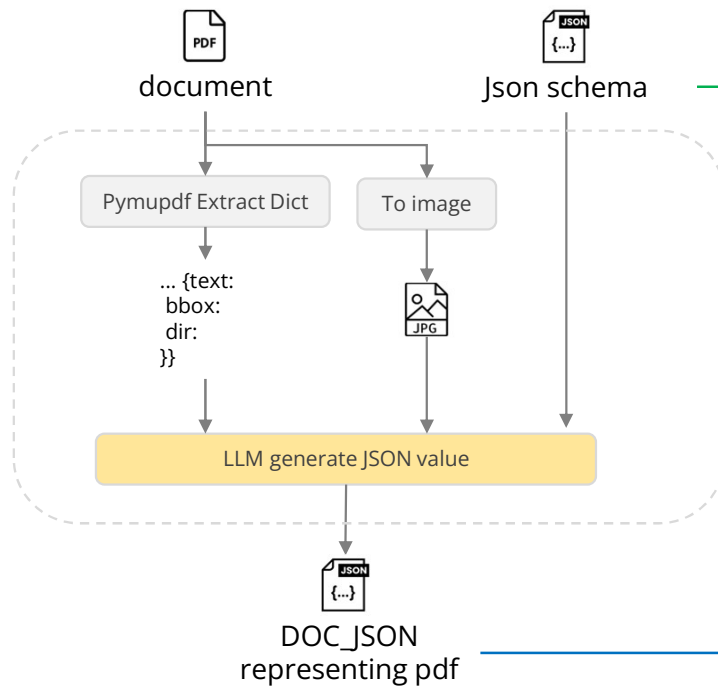
Virtual Sea Trial

Demo Architecture

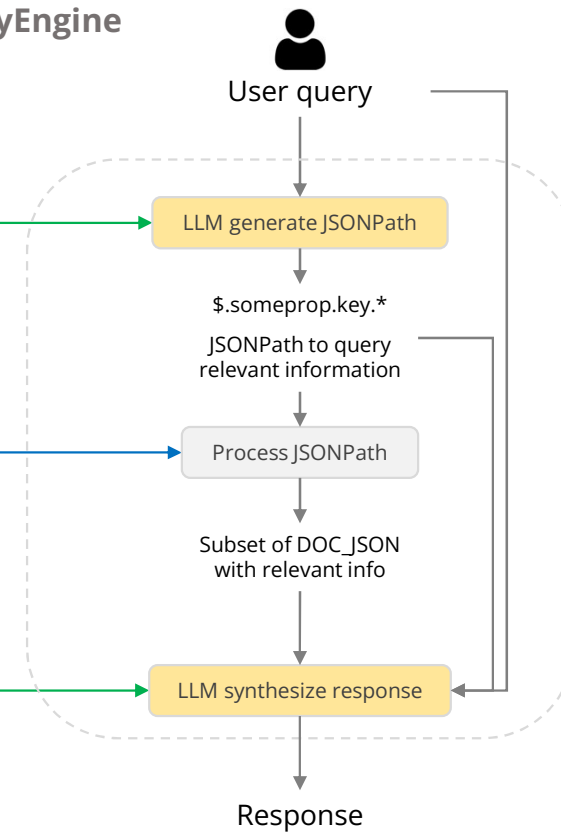
LLM USED

LLM Prompt

Pre-Processing

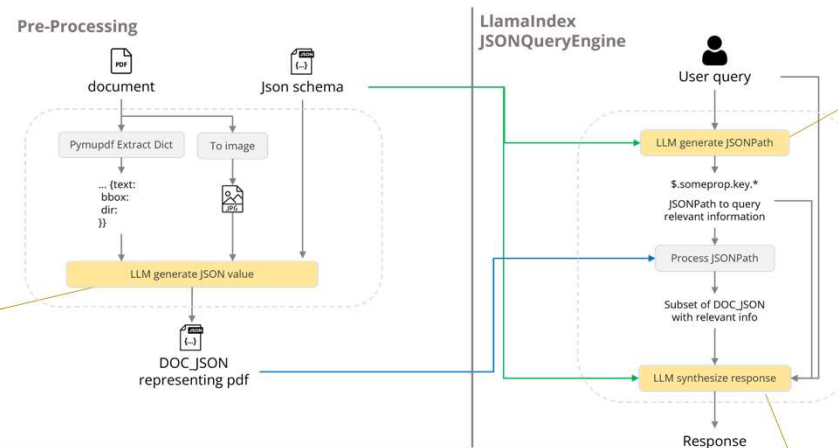


LlamaIndex JSONQueryEngine



Virtual Sea Trial

Demo Architecture w. Prompts



System Prompt: None
 User Prompt (default):
 "We have provided a JSON schema below:\n"
 "{schema}\n"
 "Given a task, respond with a JSON Path query that "
 "can retrieve data from a JSON value that matches
 the schema.\n"
 "Task: {query_str}\n"
 "JSONPath: "

System Prompt: None
 User Prompt: You are a mechanical engineer, specialized on creating, reading and understanding data sheets. You are able to extract all relevant textual and positional (as bounding boxes) data from the schematics and present them in textual format that is local and structured. The following is an image of a data sheet for a part. extract the text in it, return it as a structure json following this json schema: {json_schema_string}

To help, the following represents the pdf as a json file, containing text, position and direction information for the atomic sections of the pdf: {page_json_string}.

You separate numerical units from values into separate standard attributes but you also add a textual representation called text, where value and unit are together.

For each nested object in the JSON, provide the bounding box that outlines the property called 'text'.

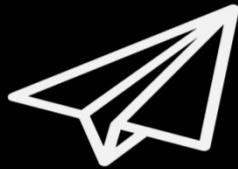
Organize it correctly using the given json schema.

Only print the data in JSON following the provided json schema, nothing else, extract all data you find.

System Prompt: None
 User Prompt:
 "Given a query, synthesize a response "
 "to satisfy the query using the JSON results. "
 "Only include details that are relevant to the query. "
 "If you don't know the answer, then say that.\n"
 "JSON Schema: {json_schema}\n"
 "JSON Path: {json_path}\n"
 "Value at path: {json_path_value}\n"
 "Query: {query_str}\n"
 "Response: "



Virtual Sea Trial



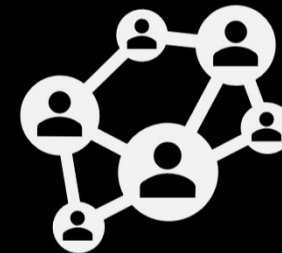
Contact

mikael.manngard@novia.fi



Web Page

virtualseatrial.fi



Social Media

#VirtualSeaTrial