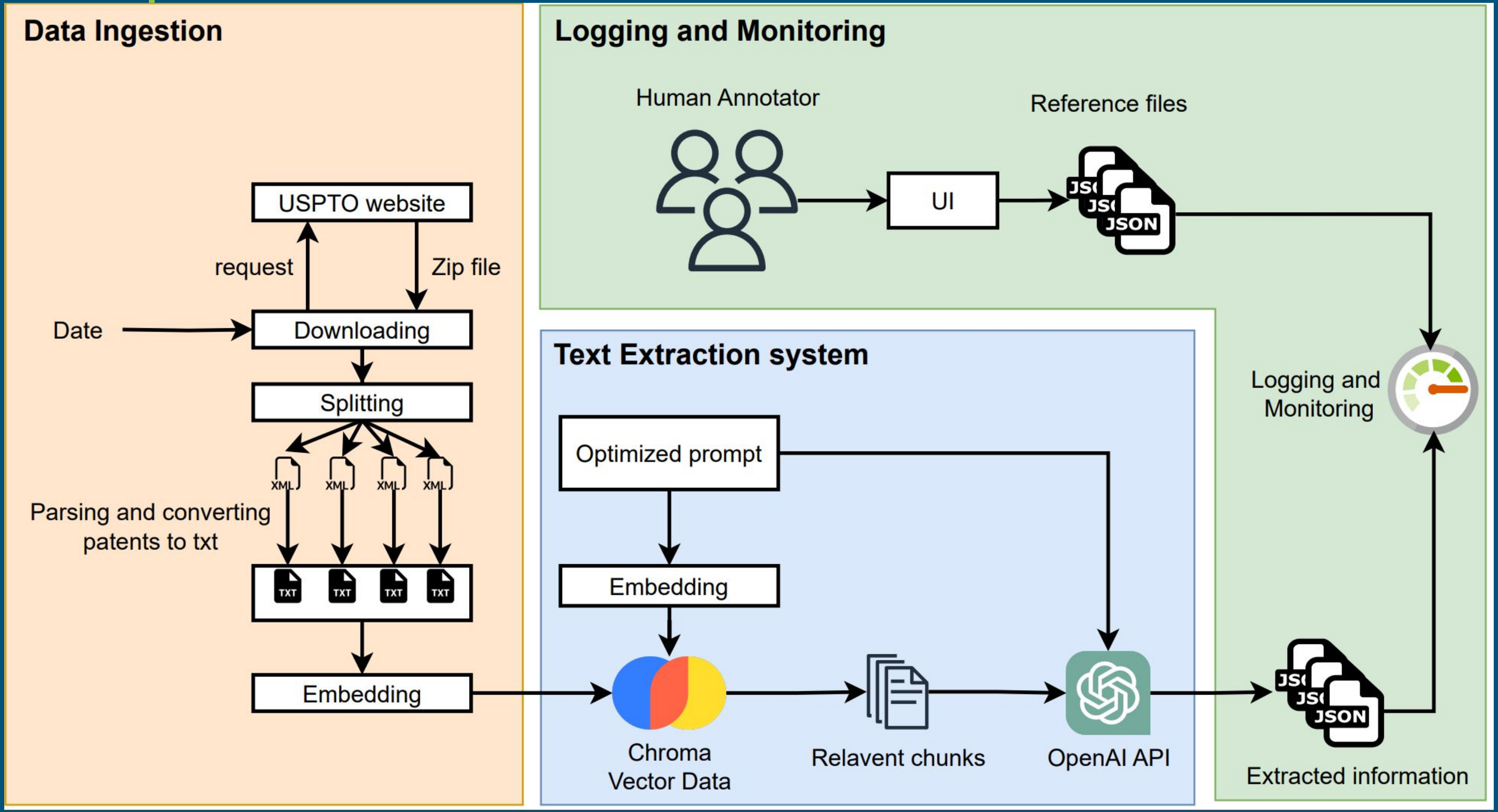
Patent analysis

By Armin Norouzi

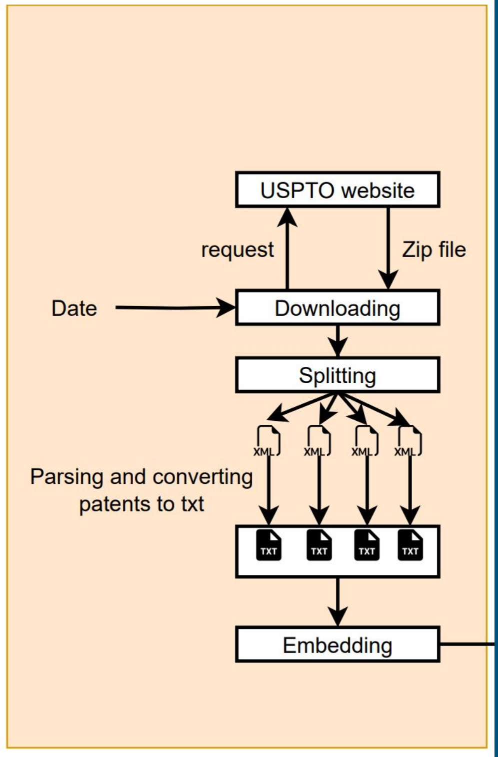
BASF

We create chemistry

System Overview

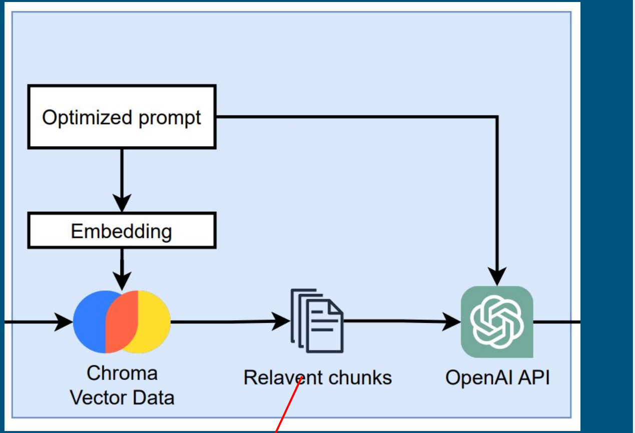


Data Ingestion



- **File Structure:** Files downloaded from *bulkdata.uspto.gov* encompass the data of patents published each week and are presented as a single concatenated XML file containing multiple patent XMLs.
- **Initial Splitting Approach:** I initially split the XML file into separate XMLs based on the ending tag `</us-patent-application>`, yielding a 78% success rate in parsing individual XMLs into text.
- **Complex File Types:** This limited success was due to the presence of two different types of XML documents within the file (`us-patent-application` and `sequence-cwu`), complicating the parsing process.
- **Improved Extraction Method:** To completely extract all patents, I revised the approach by splitting the XML file based on the `<?xml …?>` tag, allowing for more accurate separation of individual documents.

Text Extraction System

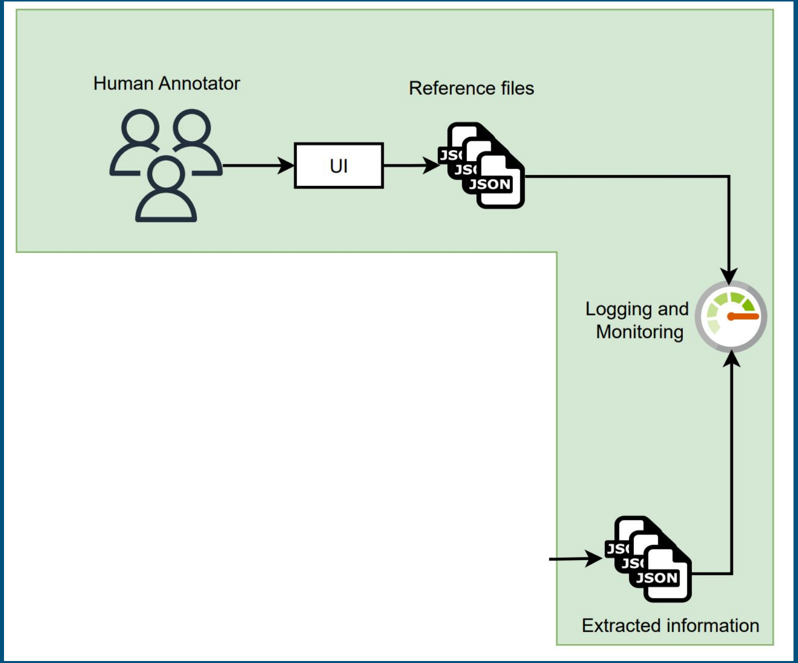


```
PROMPT_FORMAT = """
    Task: Use the following pieces of
context to answer the question at the
end.

    {context}

    Question: {question}
    """
```

```
PROMPT = """
Task: Carefully review the given patent text and extract as much physical measurements information such as
length/distance, mass/weight, time, temperature, Volume, area, speed, pressure, energy, power, electric
current
and voltage, frequency, force, acceleration, density, resistivity, magnetic field strength, and luminous
intensity as much as possible.
We are particularly interested in physical measurements including substance that was measured, Value of the
measurement, and Unit of the measurement, and measurement type mentioned in the text.
For each measurement, please provide the following details:
- The substance that was measured. (substance)
- The specific value or range that was measured. (Measured Value)
- The unit of the measurement, if provided. (Unit)
- The type of measurement being conducted (e.g., diameter, size, etc.)
Format your response in a structured JSON-like format, as follows:
{"Content": [
    {
      "Measurement_substance": "substance",
      "Measured_value": "value",
      "Measured_unit": "unit",
      "measurement_type": "type"
    },
    // ... additional measurements, if present
  ]
}
If multiple measurements are present in the text, each should be listed as a separate object within the
"Content" array.
Example: If the text includes the sentence, "The resulting BaCO3 had a crystallite size of between about 20
and 40 nm", the output should be:

{"Content": [
    {
      "Measurement_substance": "BaCO3",
      "Measured_value": "between about 20 and 40",
      "Measured_unit": "nm",
      "measurement_type": "crystallite size"
    }
  ]
}
Try to provide as complete and accurate information as possible. Print only the formatted JSON response.
"""
```

Logging and Monitoring

- **Patent Examination:** After double-checking 7 patents, it has been found that the model can successfully extract important information concerning measurements.
- **Need for Human Annotated Data:** To truly assess the model's effectiveness, there is a requirement for additional labeled data.
- **Logging and Monitoring:** Continuous logging and monitoring protocols need to be in place to ensure the proper delivery of the extracted files.
- **Quality Assurance:** Some post-processing measures will be essential to preserve the quality of the generated outputs.
- **Scheduled Updates:** The model can be configured to run a weekly task to retrieve the latest patents from the USPTO website, keeping the information current.



Results: Evaluating Outputs

```
{
            "Measurement_substance":
"substrate",
            "Measured_value": "50
rpm",
            "Measured_unit": "rpm",
            "measurement_type":
"speed"
        },
```

```
        {

"Measurement_substance": "light",
            "Measured_value": "not
provided",
            "Measured_unit": "not
provided",
            "measurement_type":
"transmitted light signal"
        },
```

```
        {

"Measurement_substance": "water",
            "Measured_value":
"6.0",
            "Measured_unit": "g",
            "measurement_type":
"volume"
        },
```

```
        {

"Measurement_substance": "hot
plate",
            "Measured_value": "80°
C.",
            "Measured_unit": "°
C.",
            "measurement_type":
"temperature"
        },
```

```
        {

"Measurement_substance": "acid
addition salt",
            "Measured_value":
"1000",
            "Measured_unit": "mg",
            "measurement_type":
"amount"
        },
```

```
        {

"Measurement_substance":
"pressure",
            "Measured_value":
"decrease with time",
            "Measured_unit": "N/A",
            "measurement_type":
"pressure change"
        },
```

Results: Initial POC vs Improvement

1. Tested Alternative options:

Name	Total Tokens	Prompt Tokens	Completion Tokens	Successful Requests	Total Cost (USD)
Kor	32615	27962	4653	32	$0.051249
Chroma	1957	1396	561	1	$0.006432
FAISS	8279	7503	776	1	$0.025613
Analyze Document Chain	17377	13733	3644	10	$0.0278875

2. Unicode and Patent ID Fix: Implemented a correction to save the patent_id and fix the unicode output by writing the output dictionary to a JSON file in UTF-8 format, ensuring that all sequences are properly displayed without any skips.

Results: Initial POC vs Improvement

3. Tested GPT-4 and compared it to GPT-3.5-turbo:

- *Performance Comparison:* GPT-3.5-turbo's performance is nearly on par with GPT-4, providing almost similar results and fair responses.
- *Runtime Efficiency:* GPT-3.5-turbo took 161.4 seconds to process 7 documents, whereas GPT-4 required 453.2 seconds, making GPT-3.5-turbo 2.81 times faster.
- *Cost Effectiveness:* GPT-3.5-turbo cost only $0.023 for processing 7 documents, as opposed to GPT-4's $0.54, making GPT-3.5-turbo 23.48 times more economical.

4. Fix error of parsing:

- *Parsing Improvement:* In the proof-of-concept (POC), the parsing success rate was 78%, but subsequent improvements have led to a 100% success rate in patent parsing.
- *Specific Filtering:* A filter has been implemented to target parents specifically from the "C" class, focusing on Chemistry and Metallurgy.

Future Works

- **Model Fine-Tuning:** Enhance precision by tuning the existing models, possibly exploring other architectures for specific patent analysis tasks.
- **Deploying Local LLM:** Implement a local Language Model using Hugging Face's Transformers library, allowing for more controlled and efficient processing.
- **Real-Time Monitoring & Analysis:** Develop real-time capabilities for immediate analysis of newly published patents, enabling quicker insights.
- **Expanding Data Sources:** Investigating and adding more sources of patent information, not limited to the USPTO, to create a more comprehensive database.
- **Enhanced Post-Processing:** Implementing advanced post-processing techniques to ensure the quality, consistency, and readiness of the extracted data.
- **Cost and Performance Optimization:** Continuously evaluate and optimize both the cost and performance of the system, considering alternative models or computing resources.