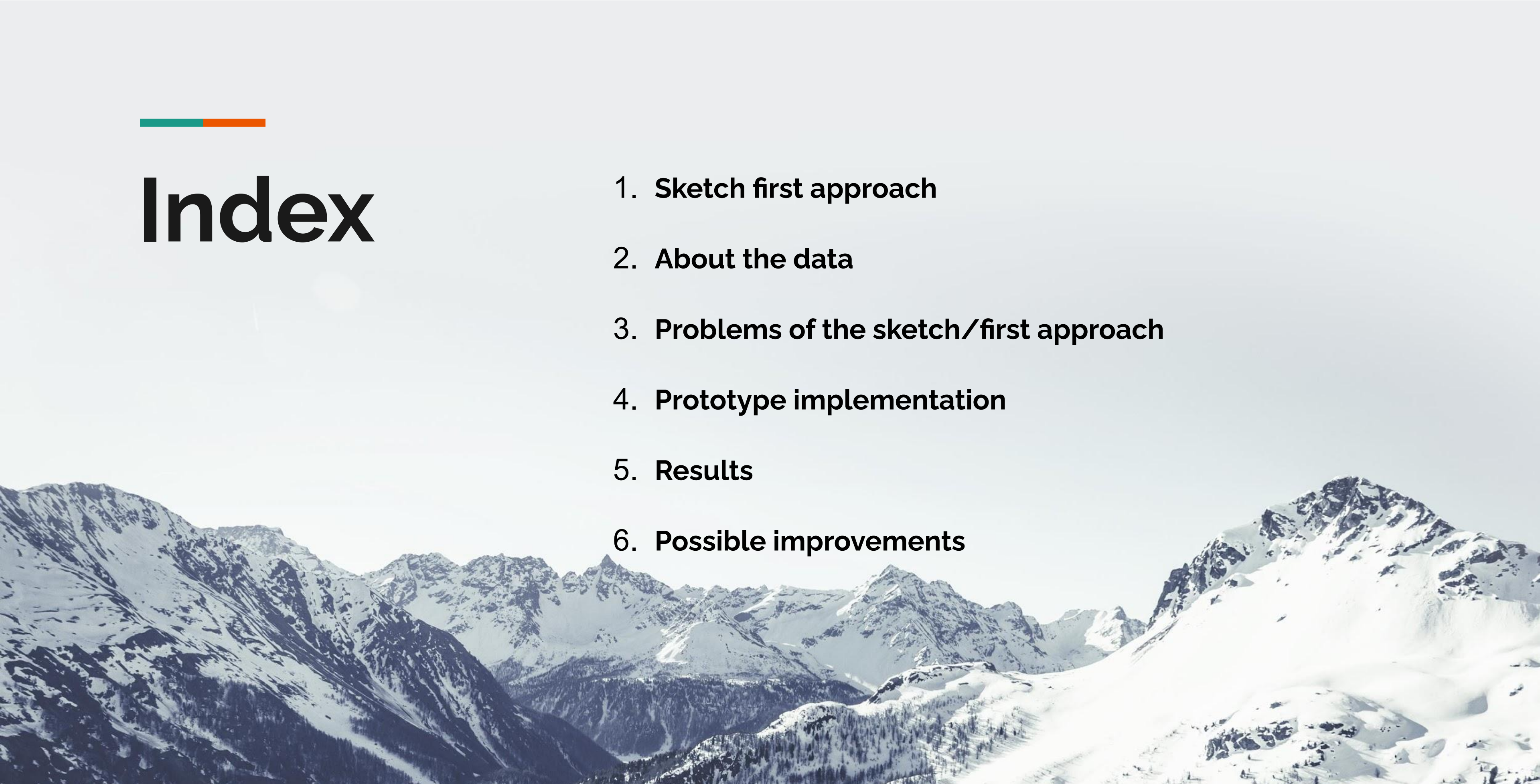
Challenge: Patent Analysis

Fernando Casabán Blasco

Index

1. **Sketch first approach**
2. **About the data**
3. **Problems of the sketch/first approach**
4. **Prototype implementation**
5. **Results**
6. **Possible improvements**



1. Sketch first approach: few-shot approach

1 **Task definition**: Extract n measurements from patents:

a. The entity
b. The property of the entity being evaluated
c. The value
d. The unit of measurement or measurement system

2 **Annotation**: Few-shot

a. Only few manually annotated examples are required
b. Represent the structure of the output.

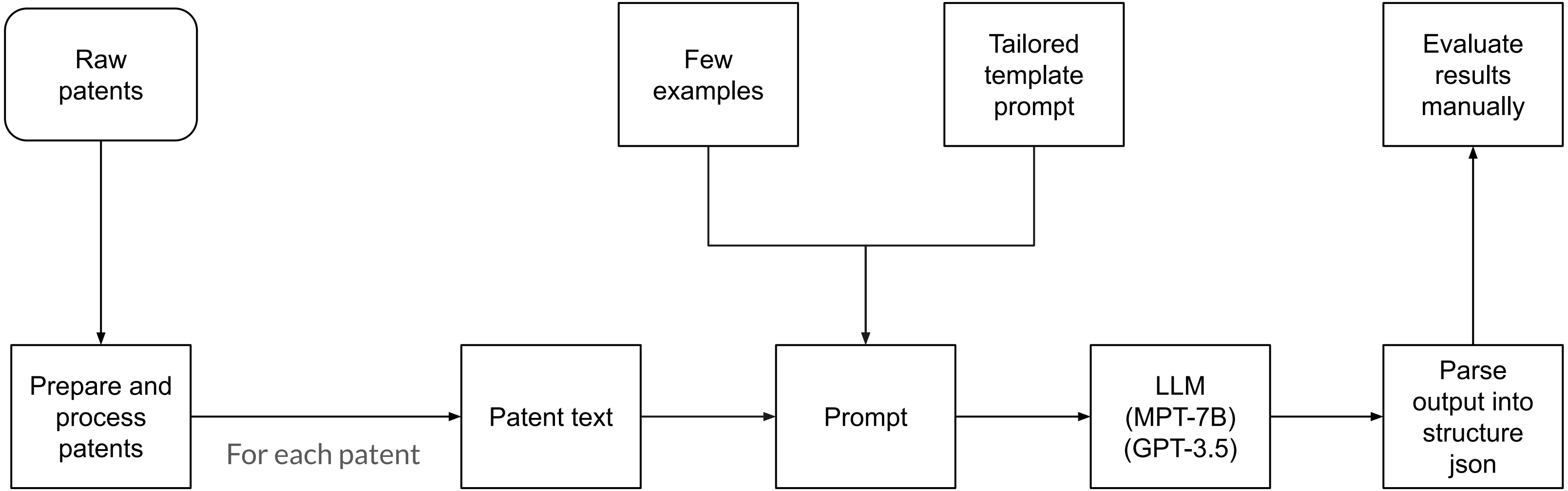
3 **Model architecture and building**:

a. Build template prompt with instructions and few examples
b. Models: MPT-7B or GPT-3.5

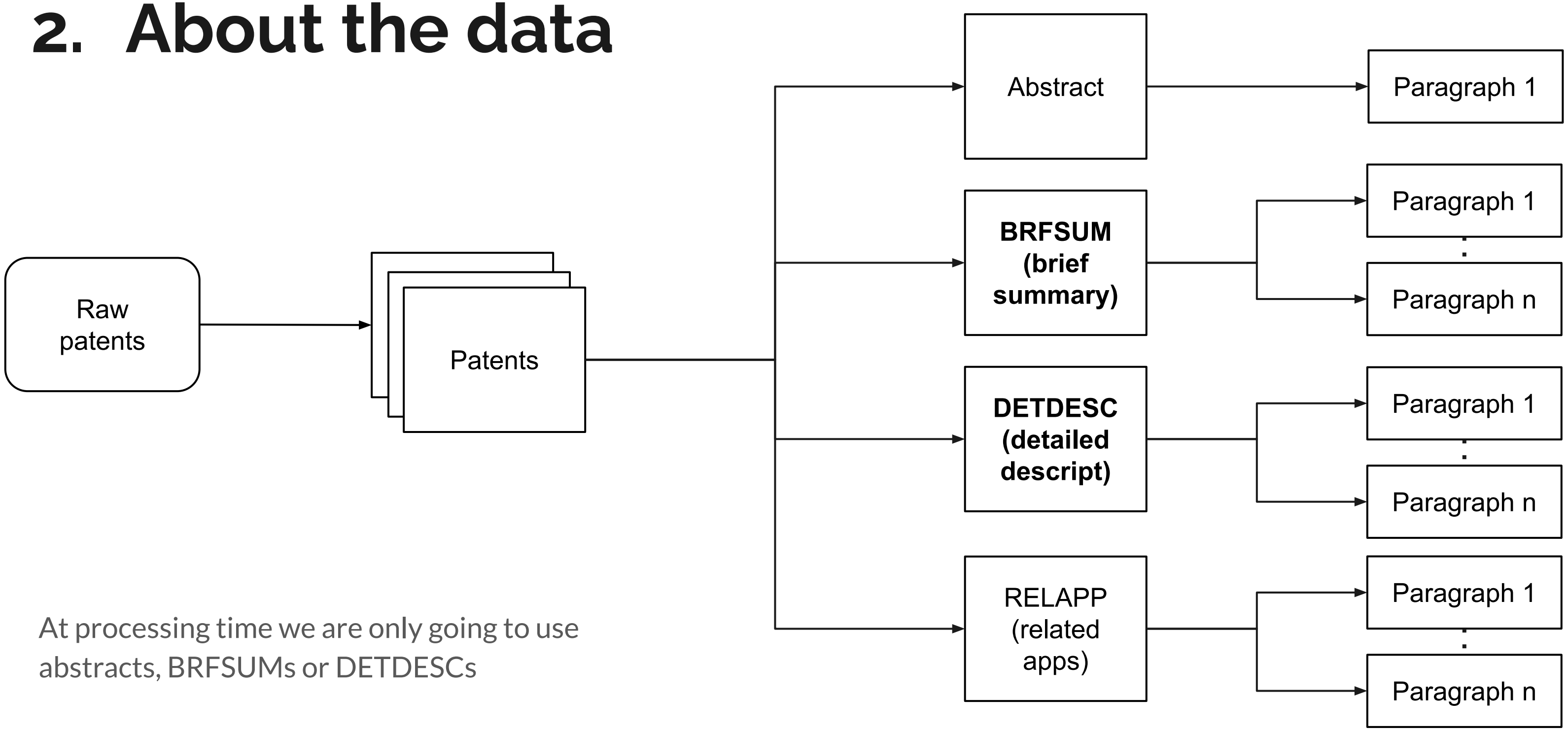
4 **Evaluation**:

a. Manual review of the model's outputs
b. Build small hold-out set

1. Sketch first approach: diagram

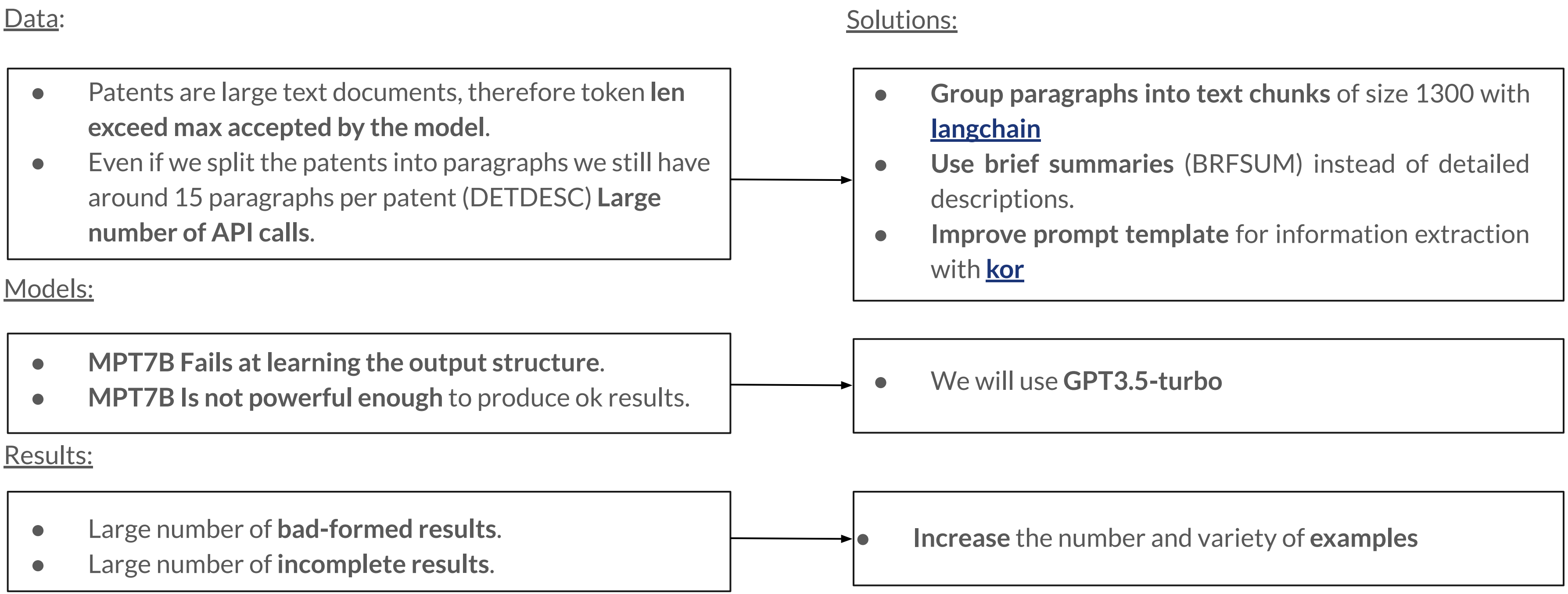


2. About the data



At processing time we are only going to use abstracts, BRFSUMs or DETDESCs

3. Problems of the sketch/first approach



Data:

- Patents are large text documents, therefore token **len exceed max accepted by the model**.
- Even if we split the patents into paragraphs we still have around 15 paragraphs per patent (DETDESC) **Large number of API calls**.

Solutions:

- **Group paragraphs into text chunks** of size 1300 with langchain
- **Use brief summaries** (BRFSUM) instead of detailed descriptions.
- **Improve prompt template** for information extraction with kor

Models:

- **MPT7B Fails at learning the output structure**.
- **MPT7B Is not powerful enough** to produce ok results.

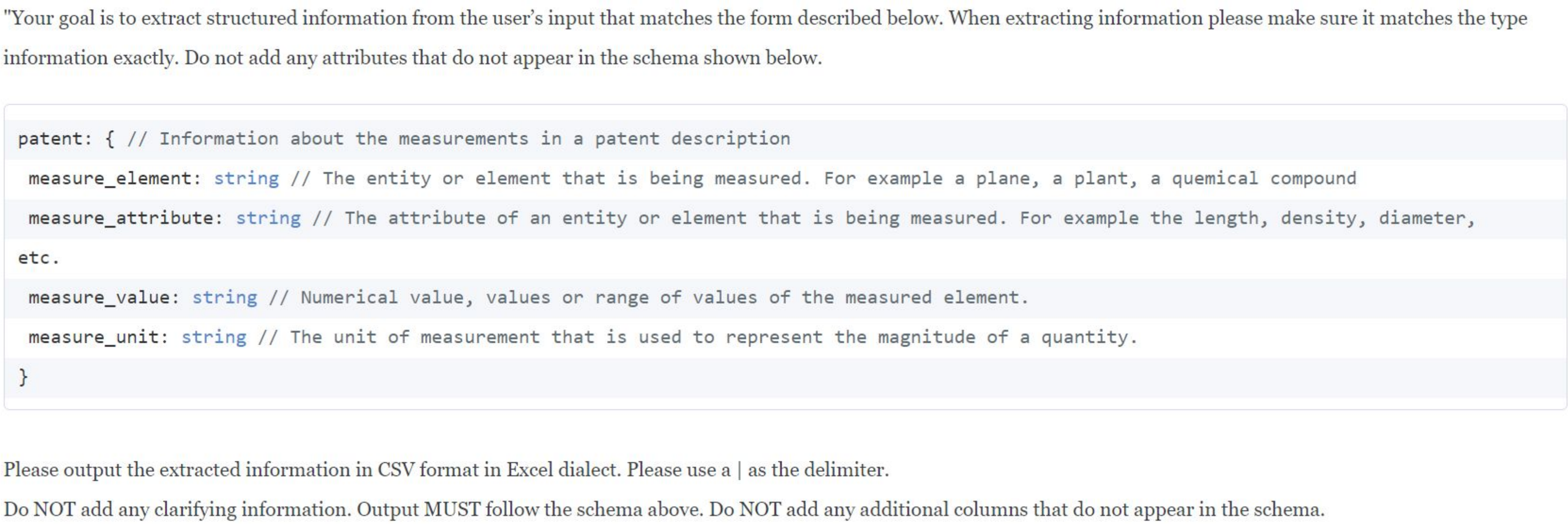
- We will use **GPT3.5-turbo**

Results:

- Large number of **bad-formed results**.
- Large number of **incomplete results**.

- **Increase** the number and variety of **examples**

4. Prototype implementation: prompt template instructions and schema



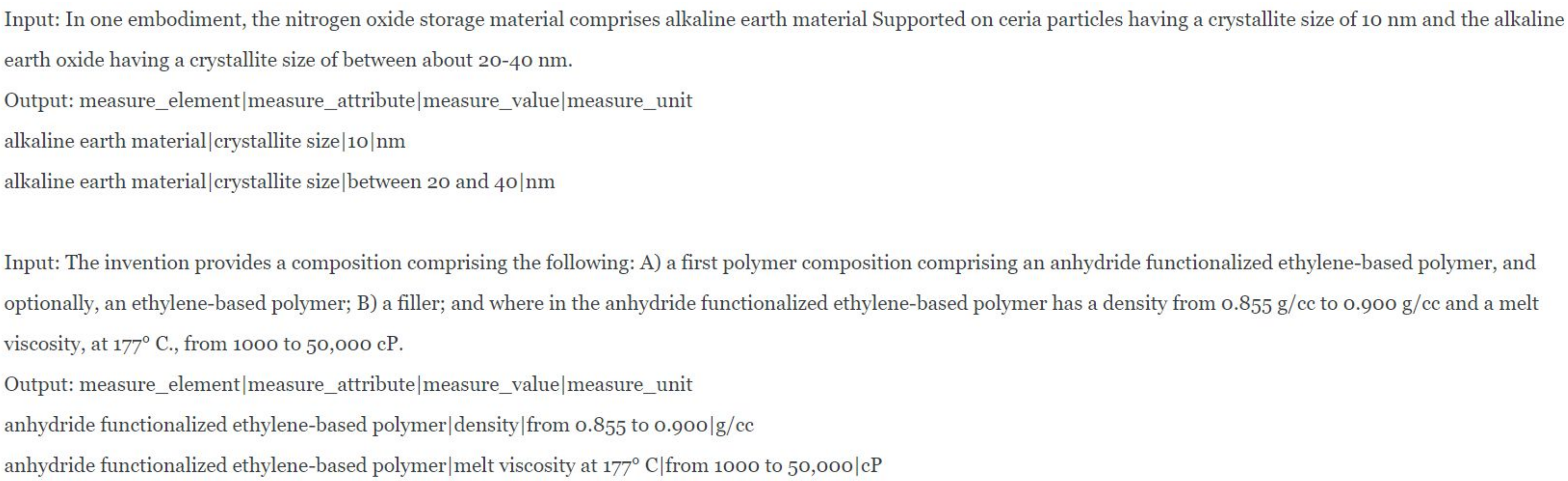
"Your goal is to extract structured information from the user's input that matches the form described below. When extracting information please make sure it matches the type information exactly. Do not add any attributes that do not appear in the schema shown below.

```
patent: { // Information about the measurements in a patent description
 measure_element: string // The entity or element that is being measured. For example a plane, a plant, a quemical compound
 measure_attribute: string // The attribute of an entity or element that is being measured. For example the length, density, diameter,
etc.
 measure_value: string // Numerical value, values or range of values of the measured element.
 measure_unit: string // The unit of measurement that is used to represent the magnitude of a quantity.
}
```

Please output the extracted information in CSV format in Excel dialect. Please use a | as the delimiter.

Do NOT add any clarifying information. Output MUST follow the schema above. Do NOT add any additional columns that do not appear in the schema.

4. Prototype implementation: prompt template few-shot examples

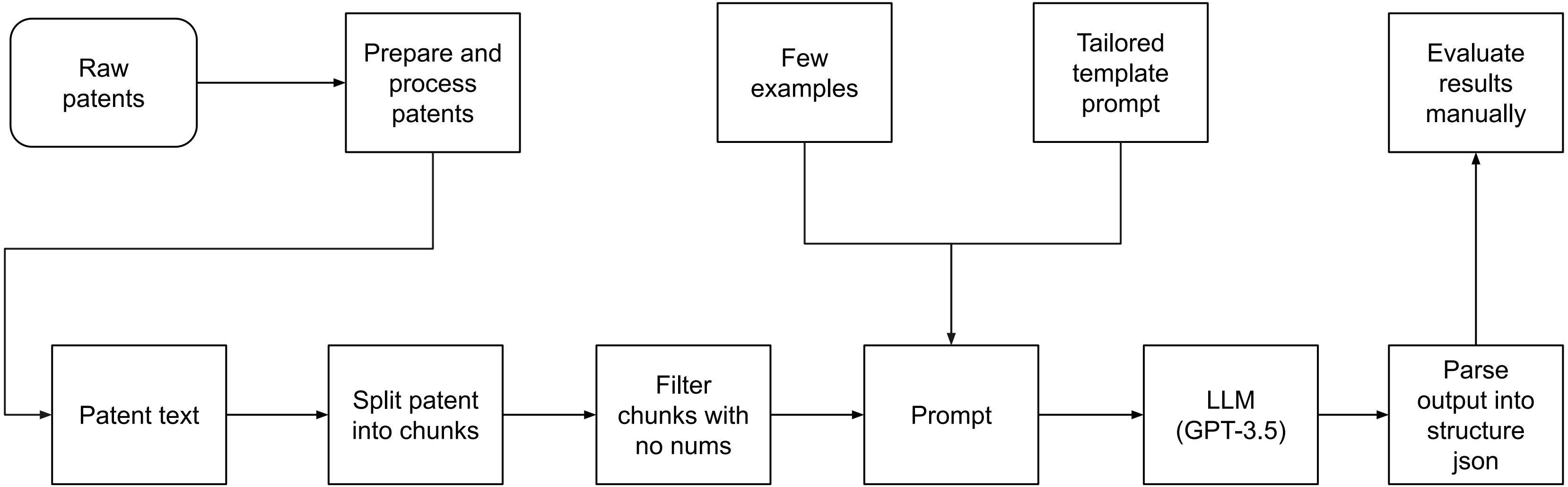


- Input: In one embodiment, the nitrogen oxide storage material comprises alkaline earth material Supported on ceria particles having a crystallite size of 10 nm and the alkaline earth oxide having a crystallite size of between about 20-40 nm.
- Output: measure_element|measure_attribute|measure_value|measure_unit
  alkaline earth material|crystallite size|10|nm
  alkaline earth material|crystallite size|between 20 and 40|nm

- Input: The invention provides a composition comprising the following: A) a first polymer composition comprising an anhydride functionalized ethylene-based polymer, and optionally, an ethylene-based polymer; B) a filler; and where in the anhydride functionalized ethylene-based polymer has a density from 0.855 g/cc to 0.900 g/cc and a melt viscosity, at 177° C., from 1000 to 50,000 cP.
- Output: measure_element|measure_attribute|measure_value|measure_unit
  anhydride functionalized ethylene-based polymer|density|from 0.855 to 0.900|g/cc
  anhydride functionalized ethylene-based polymer|melt viscosity at 177° C|from 1000 to 50,000|cP

We are using 5 examples in the final prompt template

4. Prototype implementation: diagram


Raw patents
Prepare and process patents
Few examples
Tailored template prompt
Evaluate results manually
Patent text
Split patent into chunks
Filter chunks with no nums
Prompt
LLM (GPT-3.5)
Parse output into structure json

For each patent

Reduce api calls

5. Results: discussion

Experiment on 100 patents:

- Text chunks produced: 933
- Text chunks after filtering: 431
- Measurements extracted: 1620
- Validated measures: 451

Validated measures are those that:

- Have all 4 components (element, property, value, unit).
- Have at least a num in the value field
- The unit is not N/A, unitless or similar

Conclusions:

Pros:

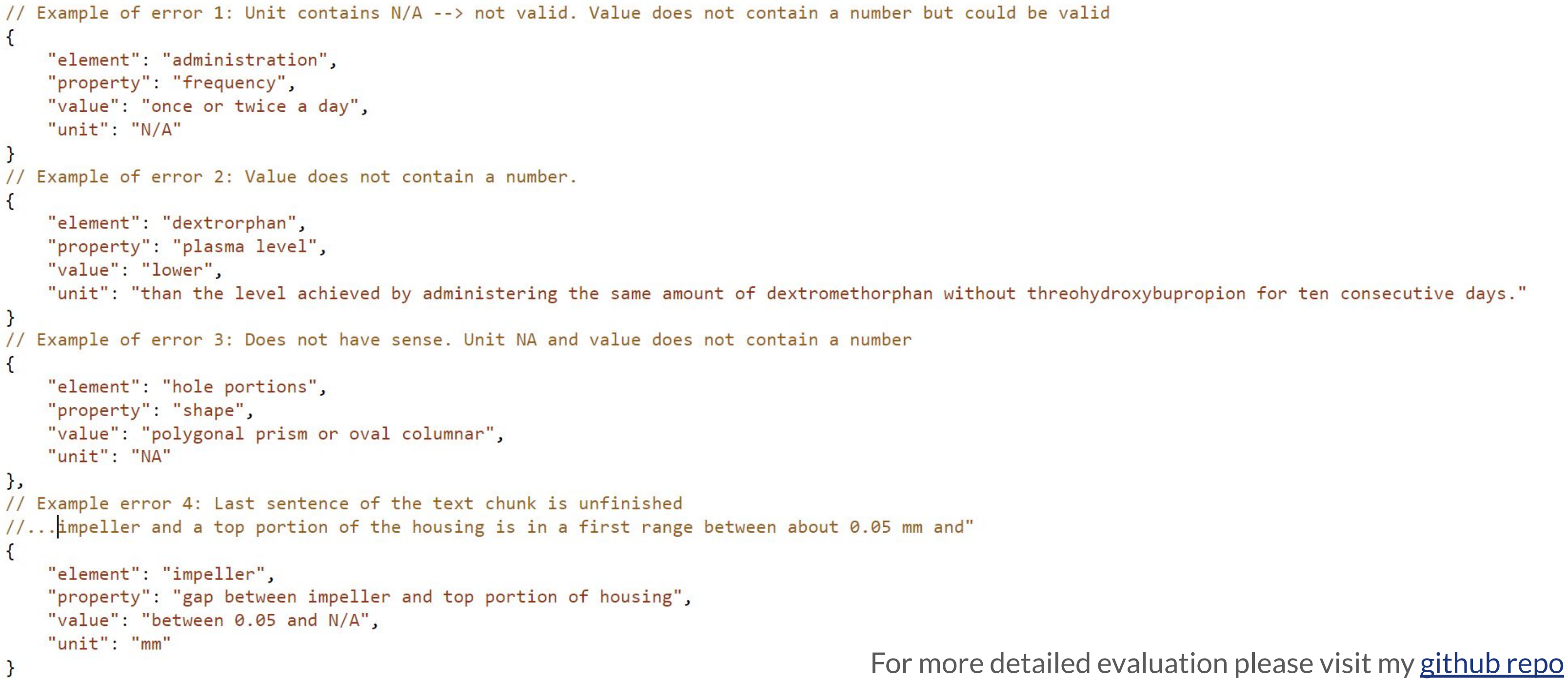
- Almost no missed measurements from text
- By filtering we avoided a lot of api calls (933 - 431 = 502)
- **Validated measurements are usually correct**

Cons:

- A lot of false positives (1620 - 451 = 1169)
- A lot of incomplete measurements (usually they are false positives)
- Even with filtering, some chunks are avoidable
- Splitting the text into chunks can break measurements

For more detailed evaluation please visit my github repo

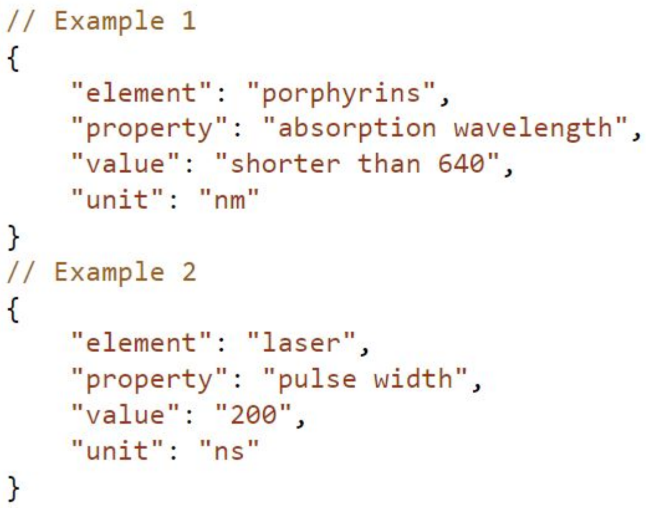
5. Results: invalid examples



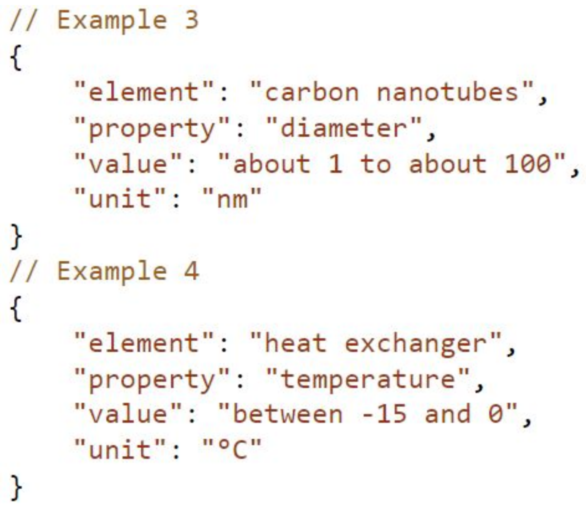
```
// Example of error 1: Unit contains N/A --> not valid. Value does not contain a number but could be valid
{
    "element": "administration",
    "property": "frequency",
    "value": "once or twice a day",
    "unit": "N/A"
}
// Example of error 2: Value does not contain a number.
{
    "element": "dextrorphan",
    "property": "plasma level",
    "value": "lower",
    "unit": "than the level achieved by administering the same amount of dextromethorphan without threohydroxybupropion for ten consecutive days."
}
// Example of error 3: Does not have sense. Unit NA and value does not contain a number
{
    "element": "hole portions",
    "property": "shape",
    "value": "polygonal prism or oval columnar",
    "unit": "NA"
},
// Example error 4: Last sentence of the text chunk is unfinished
//...impeller and a top portion of the housing is in a first range between about 0.05 mm and"
{
    "element": "impeller",
    "property": "gap between impeller and top portion of housing",
    "value": "between 0.05 and N/A",
    "unit": "mm"
}
```

For more detailed evaluation please visit my github repo

5. Results: valid examples



```
// Example 1
{
    "element": "porphyrins",
    "property": "absorption wavelength",
    "value": "shorter than 640",
    "unit": "nm"
}
// Example 2
{
    "element": "laser",
    "property": "pulse width",
    "value": "200",
    "unit": "ns"
}
```



```
// Example 3
{
    "element": "carbon nanotubes",
    "property": "diameter",
    "value": "about 1 to about 100",
    "unit": "nm"
}
// Example 4
{
    "element": "heat exchanger",
    "property": "temperature",
    "value": "between -15 and 0",
    "unit": "°C"
}
```

6. Possible improvements

➡ **Improve filtering of text chunks:**

- reduce API calls
- avoiding the exclusion of potential measurements
- Explore retrieval-augmented promotion
- Explore including vector databases

➡ **Experiment with different prompts:**

- Change template instructions
- Edit, increase or decrease provided examples
- Explore adding example with empty output to reduce noise produced by the LLM

➡ **Perform fine-tuning:**

- This will reduce costs
- The LLM would learn better the output structure
- Large annotated dataset needed
- Explore generation of synthetic data*.

➡ **Enhance the validation of extracted measurements:**

- Could miss some valids measurements
- Numbers in written form

*For an initial attempt to generate synthetic data check this file from my github repo

Thanks.