# Ingegneria dei dati
# Homework 4
# (da svolgere in gruppo)

Paolo Merialdo

# Homework 4: knowledge extraction

- From the collection of scientific papers that your team has downloaded (then, all the papers refer to the same topic), randomly select about 10 papers, each containing around 3 tables each
- The final dataset for this experiments should be composed by more than 30 tables

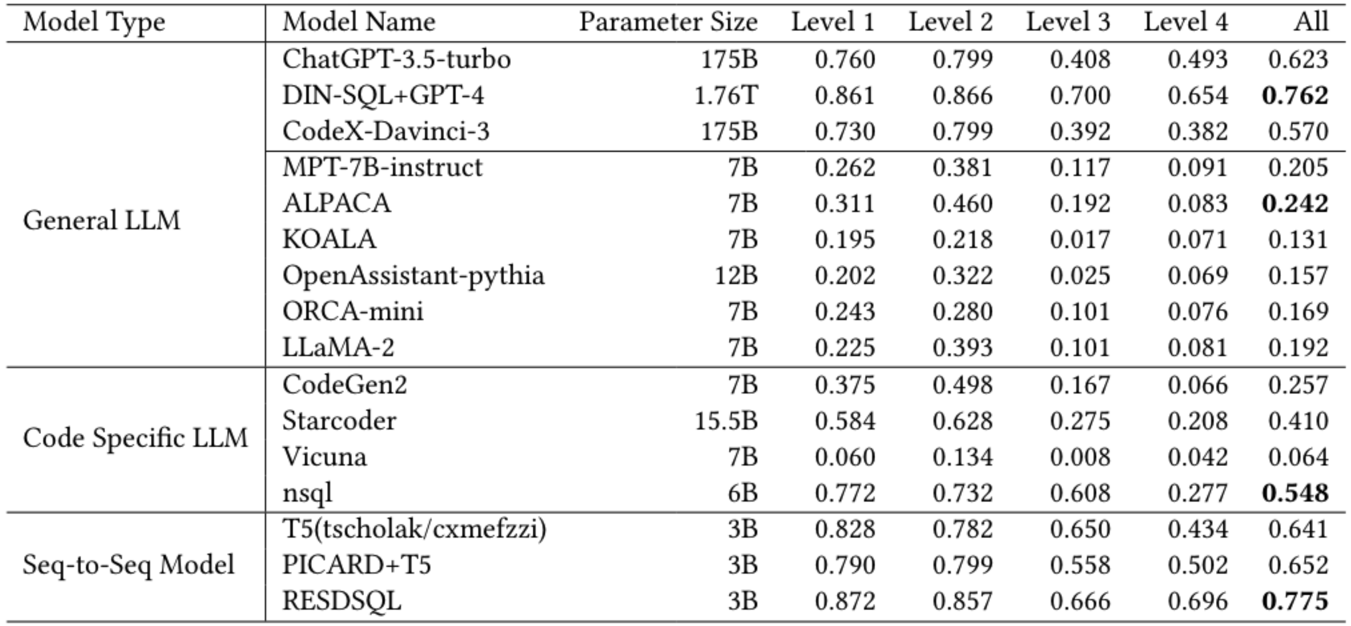
# Homework 4: knowledge extraction

- Extract the claims presented in the tables and in their associated context (references, caption, footnotes).

- Claims must be extracted according to the following format:
  - **|{Specification, Specification, ...}, Measure, Outcome|**
  - Specification: **|name, value|** pair describing the details of an experiment
    E.g.: |dataset, Spider|
    |LLM, Llama27b|
  - Measure: metric or measure used to evaluate the experiment
    E.g.: F1-measure
  - Outcome: outcome value related to metric
    E.g.: 0.89

# Example – Claims extraction (Result table)

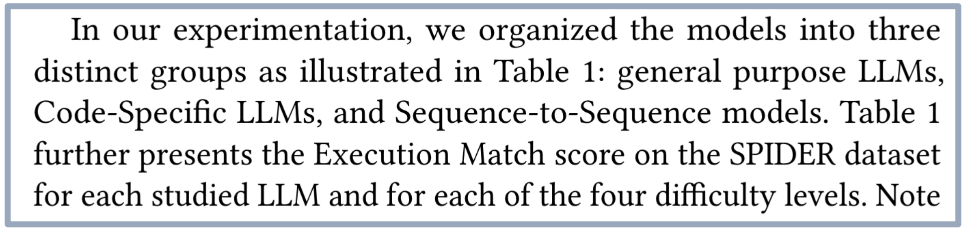
Paper: Enhancing Text-to-SQL Translation for Financial System Design
(paper cs topic: text2sql) (paper id: 2312.14725)



| Model Type | Model Name | Parameter Size | Level 1 | Level 2 | Level 3 | Level 4 | All |
|---|---|---|---|---|---|---|---|
| General LLM | ChatGPT-3.5-turbo | 175B | 0.760 | 0.799 | 0.408 | 0.493 | 0.623 |
| | DIN-SQL+GPT-4 | 1.76T | 0.861 | 0.866 | 0.700 | 0.654 | **0.762** |
| | CodeX-Davinci-3 | 175B | 0.730 | 0.799 | 0.392 | 0.382 | 0.570 |
| | MPT-7B-instruct | 7B | 0.262 | 0.381 | 0.117 | 0.091 | 0.205 |
| | ALPACA | 7B | 0.311 | 0.460 | 0.192 | 0.083 | **0.242** |
| | KOALA | 7B | 0.195 | 0.218 | 0.017 | 0.071 | 0.131 |
| | OpenAssistant-pythia | 12B | 0.202 | 0.322 | 0.025 | 0.069 | 0.157 |
| | ORCA-mini | 7B | 0.243 | 0.280 | 0.101 | 0.076 | 0.169 |
| | LLaMA-2 | 7B | 0.225 | 0.393 | 0.101 | 0.081 | 0.192 |
| Code Specific LLM | CodeGen2 | 7B | 0.375 | 0.498 | 0.167 | 0.066 | 0.257 |
| | Starcoder | 15.5B | 0.584 | 0.628 | 0.275 | 0.208 | 0.410 |
| | Vicuna | 7B | 0.060 | 0.134 | 0.008 | 0.042 | 0.064 |
| | nsql | 6B | 0.772 | 0.732 | 0.608 | 0.277 | **0.548** |
| Seq-to-Seq Model | T5(tscholak/cxmefzzi) | 3B | 0.828 | 0.782 | 0.650 | 0.434 | 0.641 |
| | PICARD+T5 | 3B | 0.790 | 0.799 | 0.558 | 0.502 | 0.652 |
| | RESDSQL | 3B | 0.872 | 0.857 | 0.666 | 0.696 | **0.775** |

**Table 1: Benchmark Results of Execution Match of all Models we tested on the "dev" SPIDER dataset**

Paragraph (reference)



In our experimentation, we organized the models into three distinct groups as illustrated in Table 1: general purpose LLMs, Code-Specific LLMs, and Sequence-to-Sequence models. Table 1 further presents the Execution Match score on the SPIDER dataset for each studied LLM and for each of the four difficulty levels. Note

1. |{|Model Type, General LLM|, |Model Name, ChatGPT-3.5-turbo|, |Parameter Size, 175B|, |Dataset, Spider dev|, |Difficulty Level, 1|}, Execution Match , 0.760|
2. ..

(claims)

# Example - Discussion

1. |{|Model Type, General LLM|, |Model Name, ChatGPT-3.5-turbo|, |Parameter Size, 175B|, |Dataset, Spider dev|, |Difficulty Level, 1|}, Execution Match , 0.760|
2. ..

- Specifications
    - |*Model type, GeneralLLM*| is located in header and index
    - |*Parameter size, 175B*| is located in header and cell
    - |*Dataset, Spider dev*| is located in caption
    - |*Difficulty Level, 1*| must be inferred from text and table header
- Measure: metric or measure used to evaluate the experiment
    - *Execution Match* metric was located caption and text, but not mentioned in the table
- Outcome: outcome value related to metric
    - Most cells of the table report outcomes but not all of them; plus, their metric is not directly mentioned in the table, but rather in the caption and the text

# Example - Claims extraction (Data table)

- Paper

Table and caption:

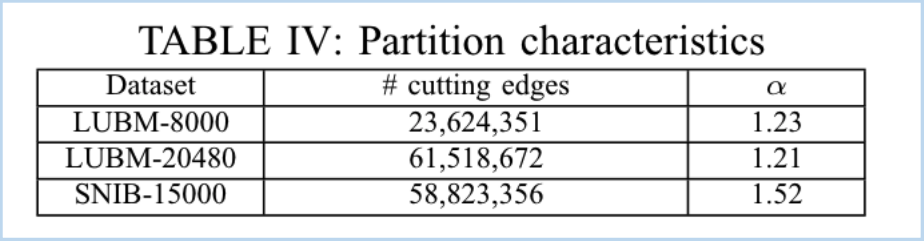
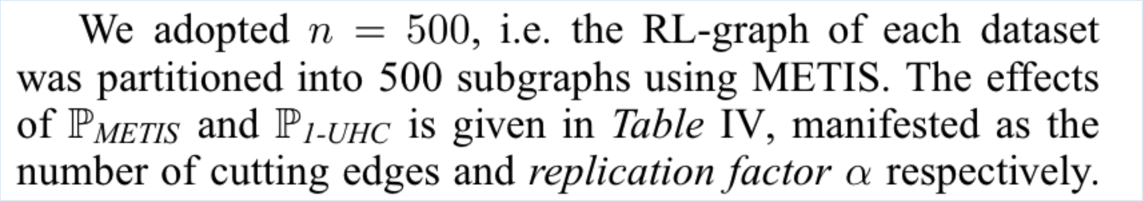


TABLE IV: Partition characteristics

| Dataset | # cutting edges | $\alpha$ |
|---|---|---|
| LUBM-8000 | 23,624,351 | 1.23 |
| LUBM-20480 | 61,518,672 | 1.21 |
| SNIB-15000 | 58,823,356 | 1.52 |

Paragraph:



We adopted $n = 500$, i.e. the RL-graph of each dataset was partitioned into 500 subgraphs using METIS. The effects of $\mathbb{P}_{METIS}$ and $\mathbb{P}_{1\text{-}UHC}$ is given in *Table* IV, manifested as the number of cutting edges and *replication factor* $\alpha$ respectively.

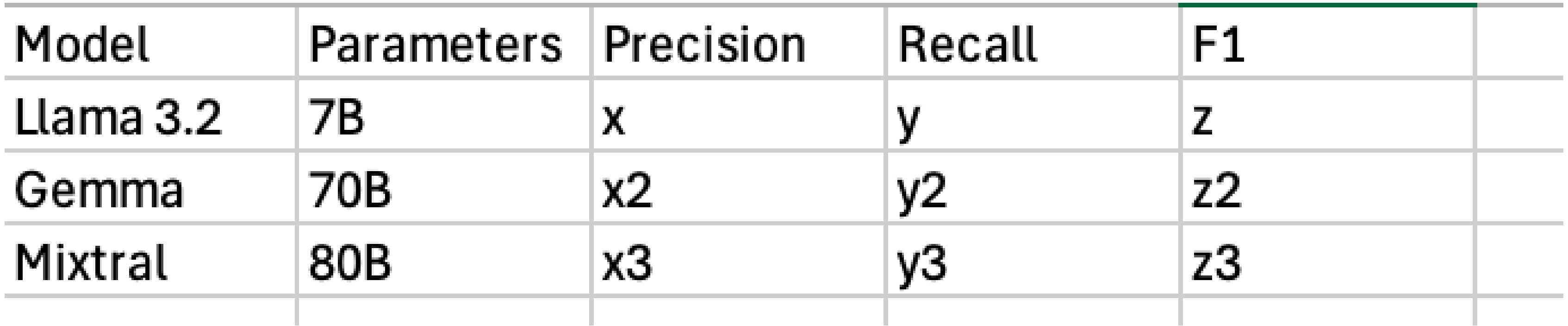
Dataset, LUBM-8000|, |RL-GRAPH partitions, 500 subgraphs|, |# cutting edges, 23,624,351|, |replication factor alpha, 1.23|}|
Dataset, LUBM-20480|, |RL-GRAPH partitions, 500 subgraphs|, |# cutting edges, 61,518,672|, |replication factor alpha, 1.21|}|
Dataset, SNIB-15000|, |RL-GRAPH partitions, 500 subgraphs|, |# cutting edges, 58,823,356|, |replication factor alpha, 1.52|}|

# Table Classes (based on structure)

- **Four** different table **classes**:
  - Relational
  - Nested relational (*tabelle relazionali nidificate*)
  - Cross-table
  - Nested cross-table (*cross-table nidificate*)

# Example Table Classes

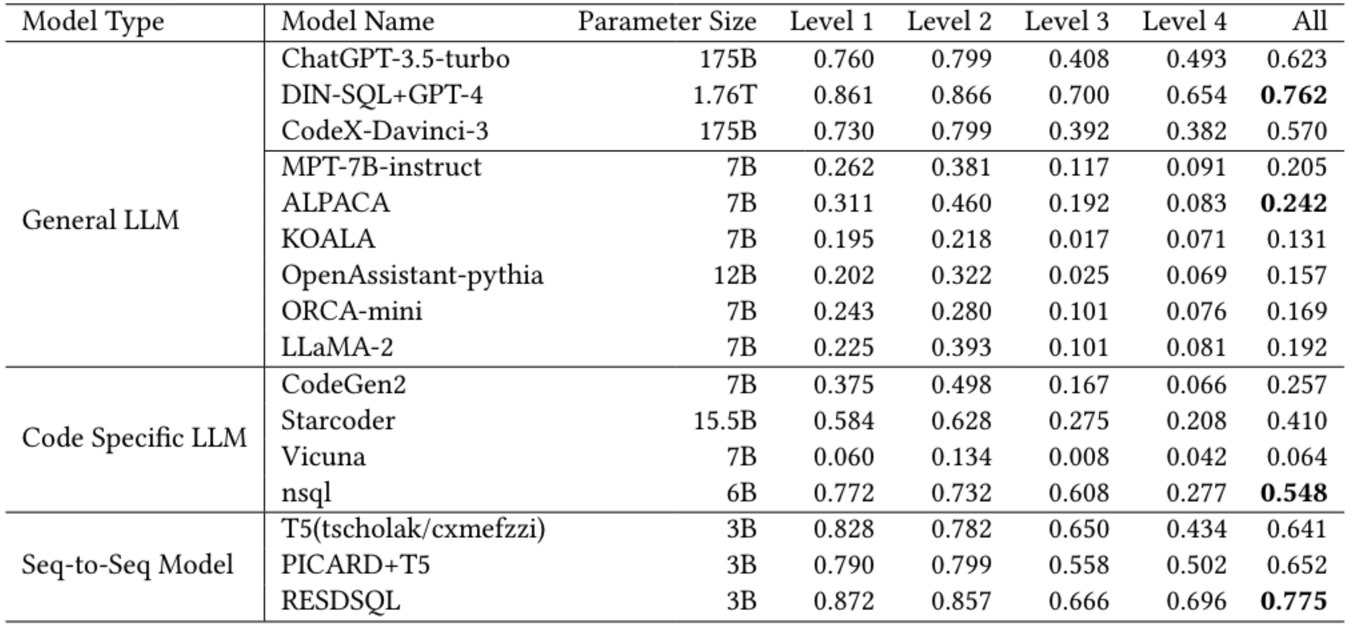
- Relational



| Model | Parameters | Precision | Recall | F1 |
|---|---|---|---|---|
| Llama 3.2 | 7B | x | y | z |
| Gemma | 70B | x2 | y2 | z2 |
| Mixtral | 80B | x3 | y3 | z3 |

# Example Table Classes

- Nested Relational

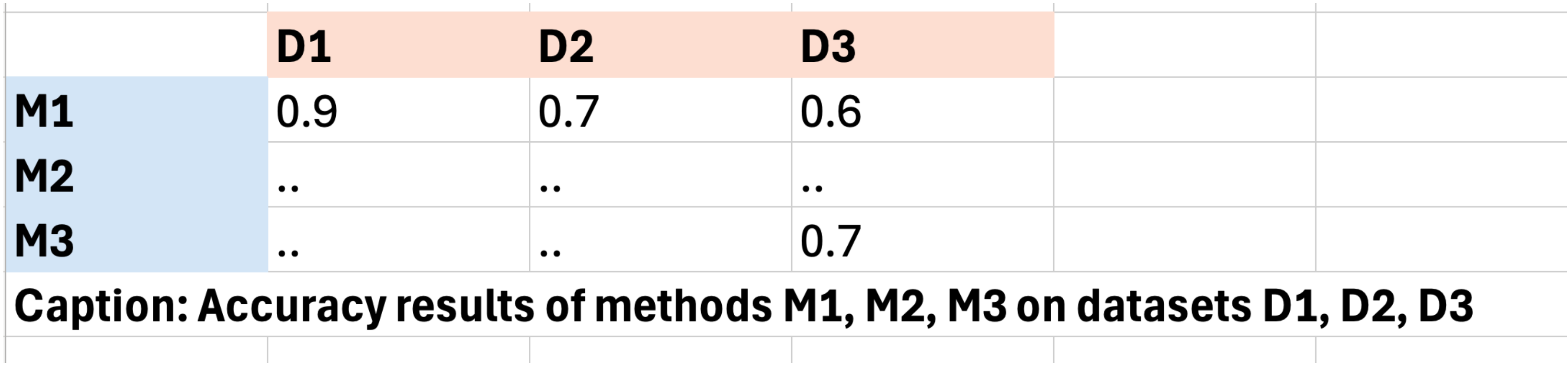


| Model Type | Model Name | Parameter Size | Level 1 | Level 2 | Level 3 | Level 4 | All |
|---|---|---|---|---|---|---|---|
| General LLM | ChatGPT-3.5-turbo | 175B | 0.760 | 0.799 | 0.408 | 0.493 | 0.623 |
| | DIN-SQL+GPT-4 | 1.76T | 0.861 | 0.866 | 0.700 | 0.654 | **0.762** |
| | CodeX-Davinci-3 | 175B | 0.730 | 0.799 | 0.392 | 0.382 | 0.570 |
| | MPT-7B-instruct | 7B | 0.262 | 0.381 | 0.117 | 0.091 | 0.205 |
| | ALPACA | 7B | 0.311 | 0.460 | 0.192 | 0.083 | **0.242** |
| | KOALA | 7B | 0.195 | 0.218 | 0.017 | 0.071 | 0.131 |
| | OpenAssistant-pythia | 12B | 0.202 | 0.322 | 0.025 | 0.069 | 0.157 |
| | ORCA-mini | 7B | 0.243 | 0.280 | 0.101 | 0.076 | 0.169 |
| | LLaMA-2 | 7B | 0.225 | 0.393 | 0.101 | 0.081 | 0.192 |
| Code Specific LLM | CodeGen2 | 7B | 0.375 | 0.498 | 0.167 | 0.066 | 0.257 |
| | Starcoder | 15.5B | 0.584 | 0.628 | 0.275 | 0.208 | 0.410 |
| | Vicuna | 7B | 0.060 | 0.134 | 0.008 | 0.042 | 0.064 |
| | nsql | 6B | 0.772 | 0.732 | 0.608 | 0.277 | **0.548** |
| Seq-to-Seq Model | T5(tscholak/cxmefzzi) | 3B | 0.828 | 0.782 | 0.650 | 0.434 | 0.641 |
| | PICARD+T5 | 3B | 0.790 | 0.799 | 0.558 | 0.502 | 0.652 |
| | RESDSQL | 3B | 0.872 | 0.857 | 0.666 | 0.696 | **0.775** |

**Table 1: Benchmark Results of Execution Match of all Models we tested on the "dev" SPIDER dataset**

# Example Table Classes

- Cross-table

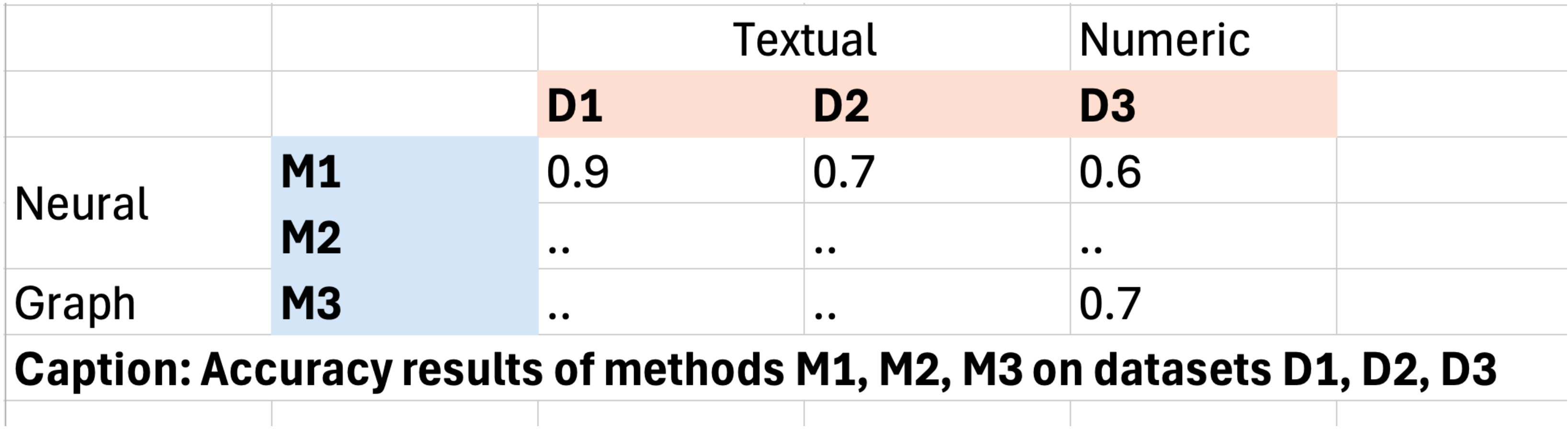


|  | D1 | D2 | D3 |
|---|---|---|---|
| **M1** | 0.9 | 0.7 | 0.6 |
| **M2** | .. | .. | .. |
| **M3** | .. | .. | 0.7 |

**Caption: Accuracy results of methods M1, M2, M3 on datasets D1, D2, D3**

# Example Table Classes

- Nested cross-table



| | | Textual | | Numeric |
|---|---|---|---|---|
| | | **D1** | **D2** | **D3** |
| Neural | **M1** | 0.9 | 0.7 | 0.6 |
| | **M2** | .. | .. | .. |
| Graph | **M3** | .. | .. | 0.7 |

**Caption: Accuracy results of methods M1, M2, M3 on datasets D1, D2, D3**

# Task 1: Claim Extraction

- Use html, caption, references and footnotes for the extraction process.
- *More specifications the better*
- Extract also a specification called |Task, …|, which represent the target task (e.g.: |task, record linkage|)

- Produce the ground truth for each table

# Task 1: File and Claims Formats

- For each pair (paper, table) you need to produce a **json file** named paperID_tableID_claims.json containing the set of the extracted claims.
    - Json file should be written following this format:
        - Claim 0: |{|Model Type, General LLM|, |Model Name, ChatGPT-3.5-turbo|, |Parameter Size, 175B|, |Dataset, Spider dev|, |Difficulty Level, 1|}, Execution Match , 0.760|
        - Claim 1: ..
    - Json format containing claims:
        - ” [
            - ‘0’: {“specifications”: {“0”: {”name”: “Model type”, “value”: “General LLM”}, “1”: {..} }, “Measure”: “Execution Match”, “Outcome”: “0.760”},
            - ‘1’: {..},
        - ]
        - **Notice that specifications are numbered!**
- ”paperID” is the id of the paper
- “tableID” is the numbered table (not the actual ID it was reported in the html)
    - First table of the paper has table_id = 1, second table has table_id=2 and so on.
    - If the paper id is “2456.7563” then the file with the associated claims from table 1 is named: “2456.7563_1_claims.json”

# Task 1: File and Claims Formatss

- Put all claims in a folder called:
  - YOUR_NAME_CLAIMS
- In which **each file** is named

  paperID_tableID_claims.json

# Task 2: Profiling

- Produce a profiling of the extracted claims.
    - Distributions of “name” in specification.
    - Distributions of “values” for each name of each specification.
    - Distributions of “ metrics”.

- Produce a spreadsheet with ColumnA key and ColumnB number of items.

- Filename should be NAME_PROFILING.CSV (or xlsx)

# Task 3: Alignment

- Align specifications names, values and metrics.
- Example:
  - In some experiments, “dataset” might be mentioned as dataset or benchmarks. Or “model” as model or algorithm.

- JSON file for the terms aligned and reproduce the profiling based on these new information.

# Task 3: Example of Alignment of Claims

- Claims to align:

1. |{|Model Type, General LLM|, |Model Name, ChatGPT-3.5-turbo|, |Parameter Size, 175B|, |Dataset, Spider dev|, |Difficulty Level, 1|}, Execution Match , 0.760|
    1. From paperid “1234.5678” table 2
2. |{|Model, SMBOP + GRAPPA|, |Dataset, Spider development set|}, Execution Match , 75.0|
    1. From paperid “6767.9898” table 4
3. |{|Model, Ours (w/ Graphix-T5)|, <Difficulty, Medium|, |Dataset, Spider|}, Execution Match , 80.7|
    1. From paper_id “3859.9017” table 1

# Task 3: Example of Alignment of Claims

| Model type | Model name | Parameter Size | Dataset | Difficulty | Metric |
|---|---|---|---|---|---|
| General LLM | ChatGPT-3.5-turbo | 175B | Spider dev | Level 1 | Execution Match |
| - | SMBOP + GRAPPA | - | Spider development set | - | Execution Match |
| - | Ours (w/ Graphix-T5) | - | Spider | Medium | Execution Match |

# Task 3: Example of Alignment of Claims

Json File:

```
{
   "aligned_names": {
      "model type": [“1234.567_2_0_0”],
      "model name": [“1234.567_2_0_1”, “6767.9898_4_0_0”, “3859.9017_1_0_0”],
..
}
 "aligned_values": {
}
```

- In “aligned_names” and “aligned_values”, for each aligned names and values you have to report as **paperID_tableID_claimID_specificationID**
- You can choose the name you prefer for the aligned value or name.

# Task 3: Filename

- Filename for the alignment is
  - YOUR_NAME_ALIGNMENT.JSON

# Termini di consegna

- Preparare un documento che descrive:
  - La soluzione usata per l'estrazione dei claim
  - Il numero di articoli e tabelle analizzate
  - La soluzione usata per valutare la correttezza dell'estrazione
- Il documento e uno zip contenente i file json per i task descritti sopra vanno consegnati entro il 10 gennaio 2025 attraverso il modulo all'indirizzo:

https://forms.office.com/e/5nmvtKgY11