# Instruct Large Language Models for Public Administration Document Information Extraction

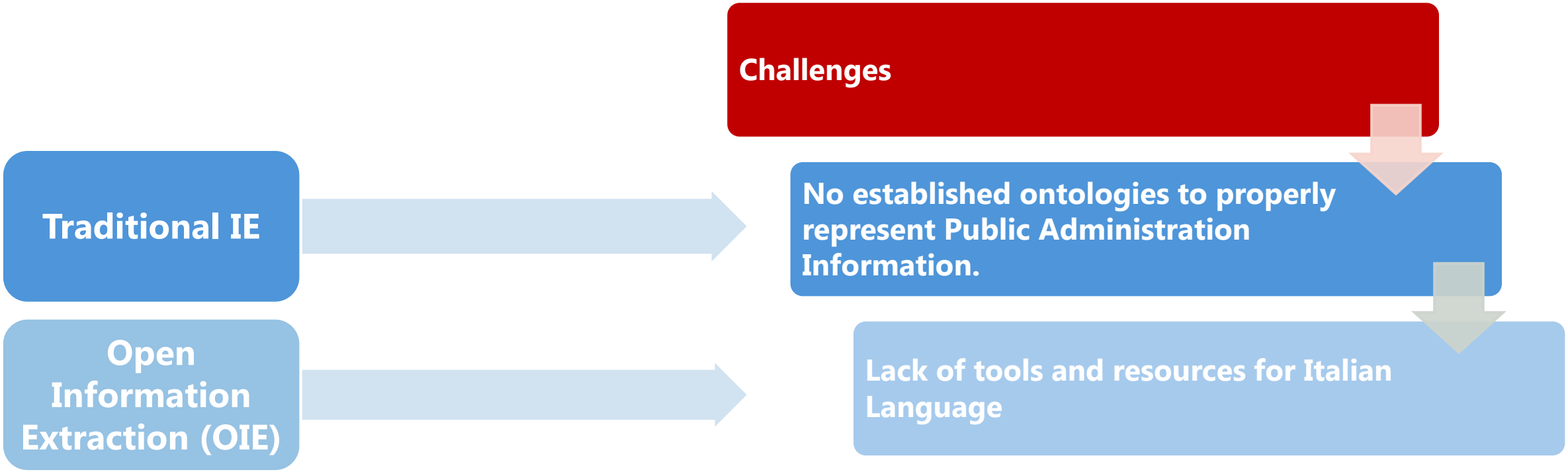
*Salvatore Carta, Alessandro Giuliani, Marco Manolo Manca, **Leonardo Piano**, Alessia Pisu and Sandro Gabriele Tiddia*

*University of Cagliari*

*Ital-IA 2024: 4th National Conference on Artificial Intelligence, May 29-30, 2024, Naples, Italy*

# Introduction

- Public Administrations (PAs) produce large volumes of disparate **unstructured data**.
- **NLP** and **Information Extraction** are valuable solutions to help PAs organize better their information.



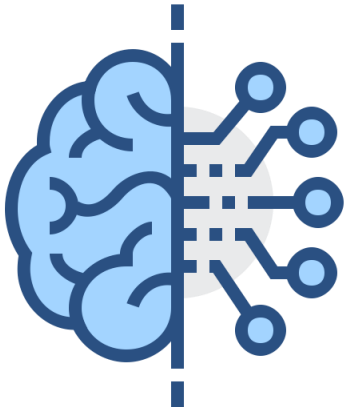
# Open Problems

Current state of the art OIE frameworks designed for the Italian language are all HandCrafted (HC).

The rules expressed by HC methods may need to be adjusted depending on the context of the PA to which they are being applied.

Neural methods might suffer less from this problem but there is no viable solution for the Italian language.

# Insight and Research Goal



LLMs  can be easily adapted to perform a new task (such as OIE), and their extensive knowledge makes them versatile in dynamic contexts.

Instruct an LLM to build a new generation of Neural OIE methods for **Italian** public administration documents.

# OIE4PA dataset

To instruct the LLM for the OIE task we leveraged the **OIE4PA** dataset (L. Siciliani, et al.)

**OIE4PA**

- Collection of triplets extracted from public tenders of the Apulia region.
- Two sets:
  - a labeled set $\mathcal{L}$ of 2000 binary triplets labeled by experts as valid or not
  - an unlabeled set $\mathcal{U}$ of 14,096 triplets extracted with WIKIOIE

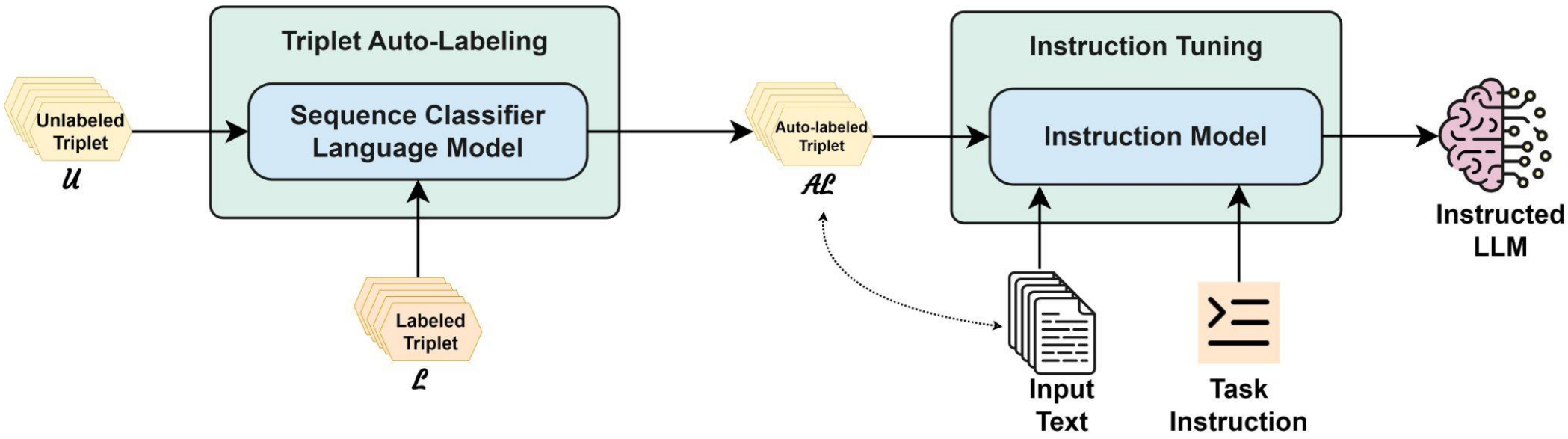
**Sentence:**

*<A norma dell'art. 51, comma 1 del Codice degli appalti, l'appalto è costituito da un unico lotto >*

**Triplet:**

<l' appalto ; è costituito da; unico lotto>

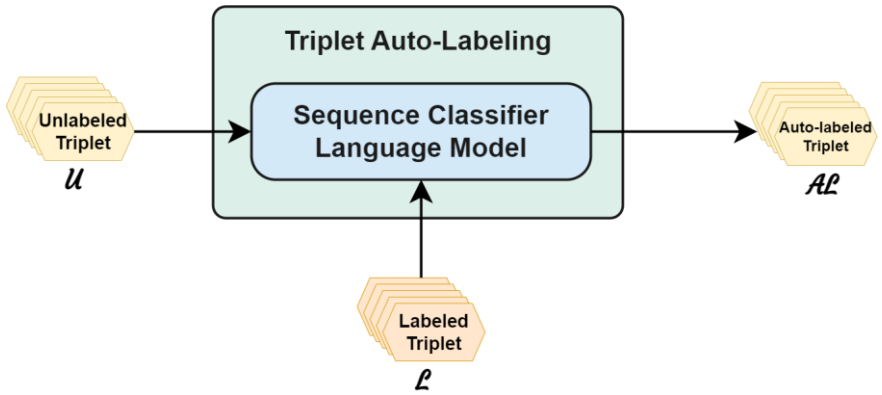
# LLM TUNING STRATEGY



# Triplet Auto-Labeling

- A Sequence classification Model is trained to identify meaningful triplets for the PA (Tenders) context.
- Triplets are concatenated into a single sentence, separating each triple member with a semicolon.
- The classifier learn to labels the triplets as valid or not.
- The trained classifier is finally employed to label the set $\mathcal{U}$, forming a new set $\mathcal{AL}$ (Auto-Labeled),

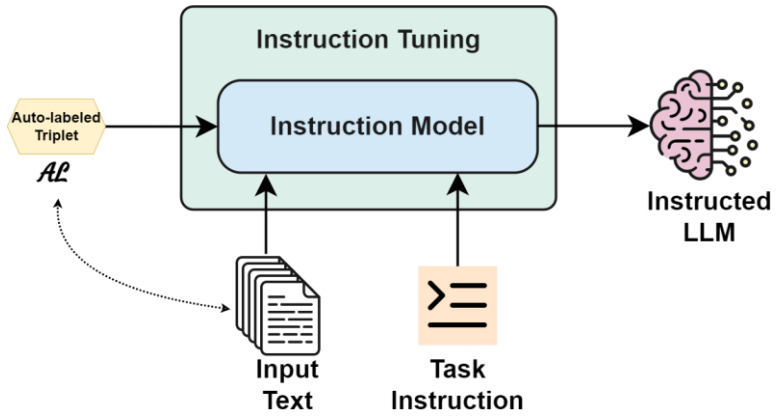
which will be exploited to Instruct the LLM on the OIE task.




<In particolare ;è richiesta ;la redazione > (0)
<L'appalto; è disciplinato da ;accordo sugli appalti pubblici> (1)
...

# Instruction Tuning

- **Instruction Tuning** involves guiding a language model through **human-like instructions** to improve its performance on a specific task.
- To instruct an **LLM** to perform **Open Information Extraction**, we transformed the $\mathcal{AL}$ triplets set into an instruction dataset, following the template: <Task Instruction, Input Text, Response>
- If a valid triplet is associated to the Input Text, the response include the valid triple otherwise we include an empty string.




**#Task Instruction:**
*<Trova quali triple semantiche esistono nel testo. Formatta l'output come [Soggetto;Predicato;Oggetto]>*

**#Input Text**:
*Il Gestore deve dichiarare nome commerciale e marca dei prodotti che si impegna ad utilizzare.*

**#Response**:
[ *Il Gestore; deve dichiarare; nome commerciale* ]

# Experimentation- Triplet Auto Labeling

- We Experimented with three Italian Language Models:
  - **Bert-Base-Italian**
  - **Italian-LegalBert**
  - **BureauBERTo**

| Model | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| Italian-Legal-Bert | 0.935 | 0.953 | 0.897 | 0.919 |
| Bert-base-Italian | 0.927 | 0.935 | 0.894 | 0.911 |
| BureauBERTo | **0.945** | **0.963** | **0.901** | **0.932** |

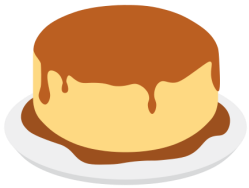
# Experimentation- Instruction Tuning

As Instruction Tuning Model, we adopted the **Flan-T5** family of models.

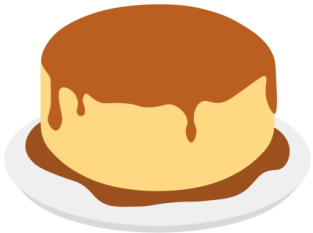
Such a choice is motivated by two reasons:

- Prior research has demonstrated the potential of such models in Information Extraction tasks.
- T5 models are **multi-lingual**, making them suitable for Italian language understanding.

We tested two different Flan-T5 sizes :



**FlanT5-xl (3b)**



**FlanT5-xxl (11b)**

| Model | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| FlanT5-xl | 0.78 | 0.74 | 0.97 | 0.84 |
| FlanT5-xxl | **0.82** | **0.78** | **0.99** | **0.87** |

# Limitations

- The model has learned to extracts a **single triple** per sentence
- The model is trained on a single dataset, limiting the model generalization abilities over multiple contexts

# Conlcusion and Future Developments

- In this work, we attempt to advance the research on neural models for Open Information Extraction (OIE) in the Italian language, applying such a technology to the PA context, particularly public tenders.
- Neural  methods heavily depend on the quality and availability of datasets, highlighting the need to develop additional datasets to compare and improve future approaches.
- Our work provides a pioneering model for comparison contributing to the existing literature.
- As for Future development, we plan to create new datasets with consistent document-level annotations  to train the model to understand relationships that may exist between different sentences.

# Thanks For Your Attention!

leonardo.piano@unica.it