MISTRAL 7B
FINETUNING
For Automatic Knowledge Graph Contruction

Davide Giardini - 897473

TABLE OF CONTENTS —

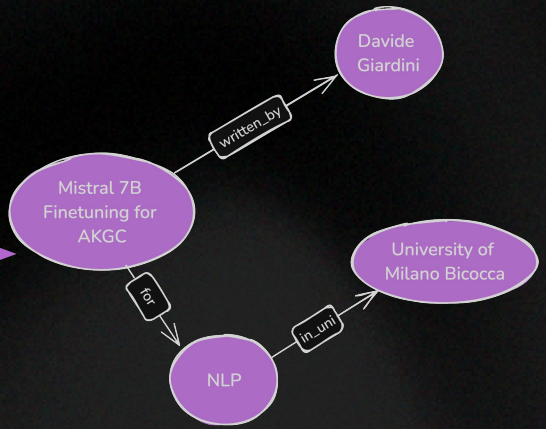
01 OBJECTIVES

01
OBJECTIVES —

Introduction to Automatic Knowledge Graph Construction

The project “Mistral 7B finetuning for AKGC” was written by Davide Giardini for the course of NLP in University of Milano Bicocca.

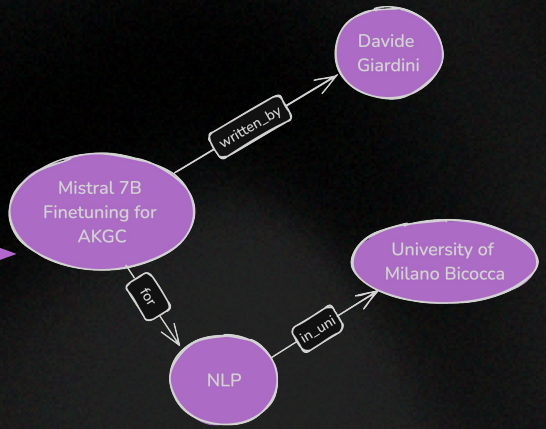
- **Named Entity Recognition**
- **Named Entity Linking**
- **Coreference Resolution**
- **Relation Extraction**



01
OBJECTIVES —

Main Idea

The project “Mistral 7B finetuning for AKGC” was written by Davide Giardini for the course of NLP in University of Milano Bicocca.



01
OBJECTIVES —

Problem Statement

This work does not aim to replace the pipeline in its entirety.
Rather, we want to evaluate the performances of finetuned LLMs on a closed environment with specific limitations:

Predefined Schema	Specify to the LLM the entities and relationships to search for. This also resolves the problem of **consistency**.
Single Documents	The extracted triples should be consistent and follow the specified KG schema, but the model is not asked to link them to the entities found in other documents.
No node properties	No node's properties will be required to be extracted from the documents, only node's labels and relationship's types.
Short text inputs	We are going to deal with short text inputs, derived from a graph of maximum 12 nodes.

01
OBJECTIVES —

Goals

Provide the community with an easy-to-implement, computationally inexpensive tool to extract triples from a text. This is done to replace the current prompt-based methods with a new system that is hopefully more accurate. Moreover, unlike most of the other tools, we aim at empowering the user with the ability of specifying the structure of the KG.

Evaluate the effectiveness of LLMs fine-tuning on a synthetic dataset on a restricted AKGC task, in order to assess whether more research should be developed in this direction.

Identifying which problems arise with the implementation of a LLM on the task of AKGC, and offer insights into their potential solutions for future research.

02

DEVELOPMENT

02 DEVELOPMENT —

Subgraphs Extraction

1.	Recommendations	Train & Test
2.	Legis-graph	Train & Test
3.	Recipes	Train & Test
4.	Listings	Train & Test
5.	Graph-data-science	Test
6.	wwc2019	Test

- Select One Random Node
- Select Randomly the size of the graph (6-12 nodes)
- Select randomly the depth of the graph (2-6)
- Extract the subgraph with the selected parameters

From the 4 databases used for Train

- 2400 (600 each) used for training
- 400 (100 each) used for validation
- 400 (100 each) used for test

From the 2 databases used for Test

- 200 (100 each) used for test

02
DEVELOPMENT —

Text Generation

Imagine being a text generator from Knowledge Graphs.
Based on the triples provided in the context, generate a short text containing all the information contained in the triples. Make sure not to add any information of the entities mentioned in the triples that is not coming from the knowledge graph. Even though the usage of pronouns is allowed, make sure not to modify the names of the entities.
The text you generate should not be a simple mention of all the facts stored in the triples, but you should write them in an original way. The text should resemble a $style.

This is the KB structure:
$KB_structure

Context:
$context

Structure of the current KB. It is inserted in the prompt in order to provide more information on what the triples of the context represent.

Extracted Triples

- Blog Article
- Wikipedia Article
- Newspaper Article
- Reddit Post
- YouTube Script
- Podcast Script

02
DEVELOPMENT —

Text Formatting

<s>[INST]Imagine being a Knowledge Graph constructor from unstructured text.
Following the schema provided, extract all the triples you can find in the text.

Schema:
$KB_structure

Context:
$Text[/INST]

Extracted Triples:
$Triples</s>

02
DEVELOPMENT —

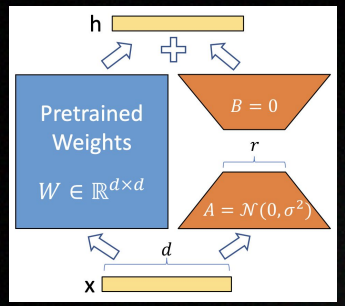
Finetuning: LoRA and QLoRA

Aghajanyan et al.
(2020)

Pre-trained language models have a low “intrinsic dimension” and can still learn efficiently despite a random projection to a smaller subspace.

LoRA

Edward J Hu et al.
(2021)



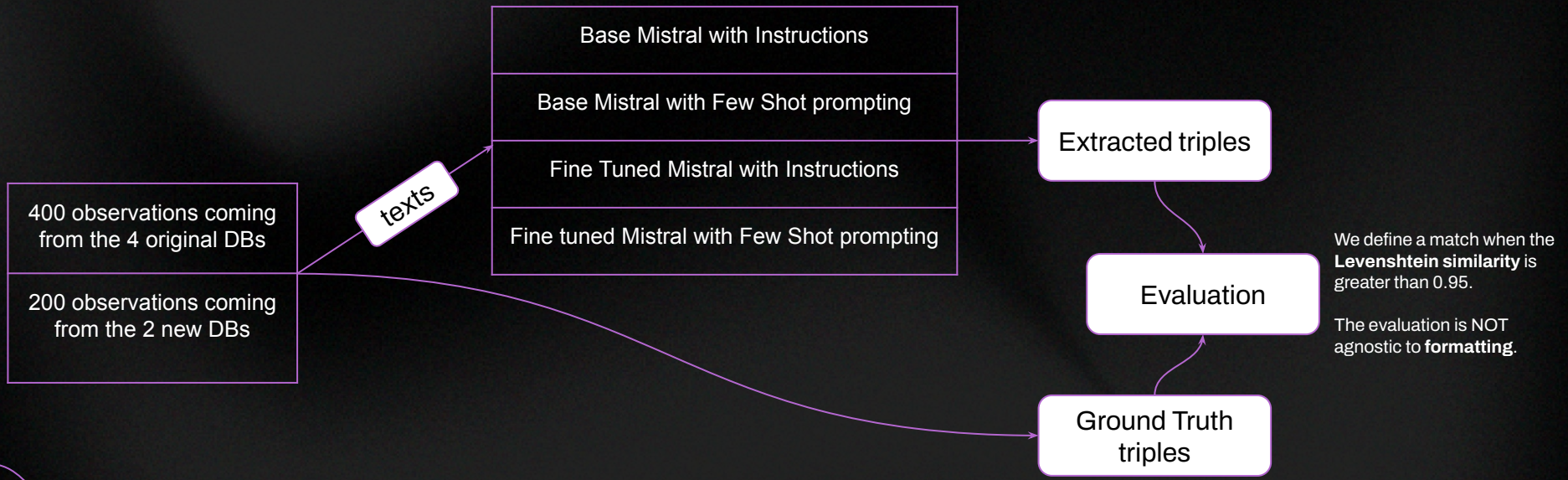
QLoRA

Tim Dettmers et al.
(2024)

- Paged Optimizer
- NormalFloat
- Double Quantization

02
DEVELOPMENT —

Inference and Evaluation



03
RESULTS

RESULTS

4 original DBs

Method	Average Precision	Average Recall
Base Model with Instructions	0.25	0.24
Base Model with Few Shot prompting	0.63	0.54
Fine Tuned model	0.81	0.77

RESULTS

2 DBs reserved for Testing

Method	Average Precision	Average Recall
Base Model with Instructions	0.09	0.08
Base Model with Few Shot prompting	0.72	0.57
Fine Tuned model	0.37	0.31
Fine Tuned model with Few Shot prompting	0.69	0.55

04
FineTuning for Few-Shot prompting

04

FT for FS prompting

Text Formatting

Method	Average Precision	Average Recall
Base Model	72.13	56.71
Fine Tuned (before)	68.85	55.06
Fine Tuned (50 steps)	71.82	66.41
Fine Tuned (150 steps)	67.99	61.96
Fine Tuned (300 steps)	51.35	45.34

<s>[INST]Imagine being a Knowledge Graph constructor from unstructured text.
Following the schema provided, extract all the triples you can find in the text.

Schema:
$KB_structure

Here are some examples:
Context:
$example1_text
Extracted Triples:
$example1_triples

Context:
$example2_text
Extracted Triples:
$example2_triples

Context:
$example3_text
Extracted Triples:
$example3_text

Context:
$Text [/INST]
Extracted Triples:
$Triples </s>

THANK YOU

Davide Giardini - 897473

MISTRAL 7B
FINETUNING
For Automatic Knowledge Graph Contruction