SAPIENZA
UNIVERSITÀ DI ROMA

Homework Presentations

MNLP Course 2024/2025

Group:
Giuminia

Group Member:
Giulia Alemanno

Homework 1: Multiclass Classification of Cultural Items

What makes something cultural agnostic, cultural representative or cultural exclusive?

Bread

⬇

CULTURALLY AGNOSTIC

Colosseum

⬇

CULTURALLY REPRESENTATIVE

Caponata

⬇

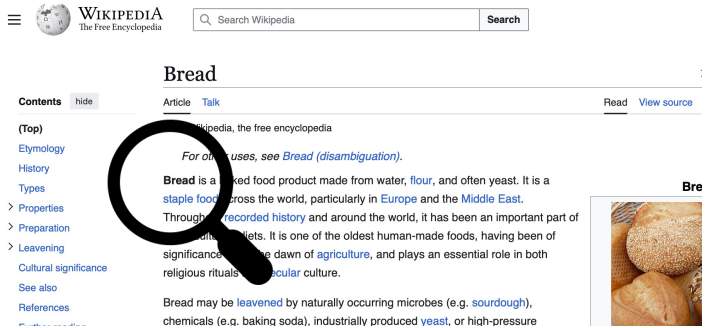
CULTURALLY EXCLUSIVE

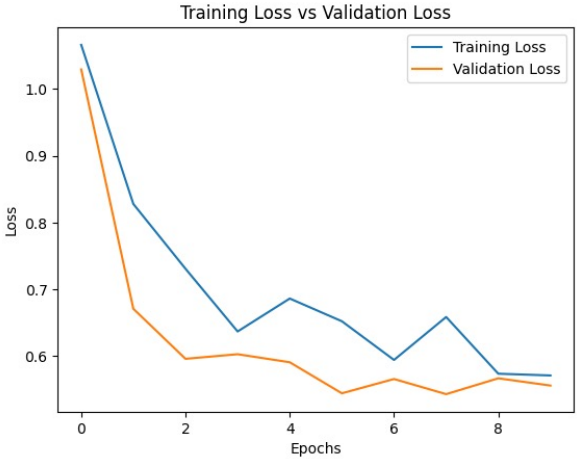
Task 1: LLM based solution

Finetuning a Transformer-based model: roBERTa base (ecoder-only)

For each item, listed in the dataset given, we extracted the first 100 characters found in its English Wikipedia page's introduction

The introduction and each item's description were tokenized and given as input to the model



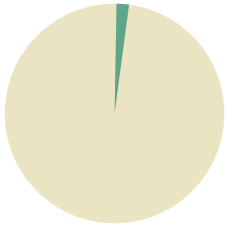


Accuracy	Precision	Recall	F1 Score
0.80	0.79	0.78	0.79

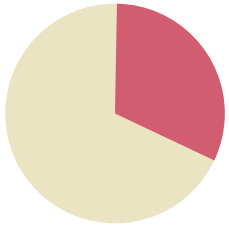
Task 2: non-LLM based solution

Feature Selection: Failed attempts

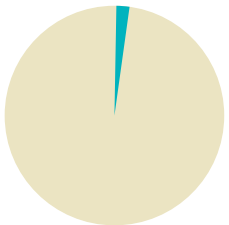
- **Native Label (P1559, P1148)**



1.7% Culturally Agnostic



32% Culturally Representative



1.8% Culturally Exclusive

Interesting property in the cultural context but it is not always present in the data

- **Sources**

Should all sources listed for a item be treated equally?

It is necessary to define **RELIABLE** sources in order to use this criteria for classification

Wikipedia definition of a source: *«Something as reliable as the context in which it is referenced in»*

Final Features Chosen

Is_human

Boolean property checking if the entity is a human

Has_country

Boolean property checking if the entity has the «country» property

Countries_found

Required the definition of a csv containing all countries and their denonyms

Countries found via WordToVec in the introduction of the English Wikipedia page for the item

Inlink_count

Centrality in the wikidata graph: counts each item's connections

Mean of connections per class:
- CA 15000
-CR 6800
- CE 170

Num_wikilinks

Number of Wikipedia pages in other languages made for the item

Num_properties

Counts all the properties listed in the item's wikidata page

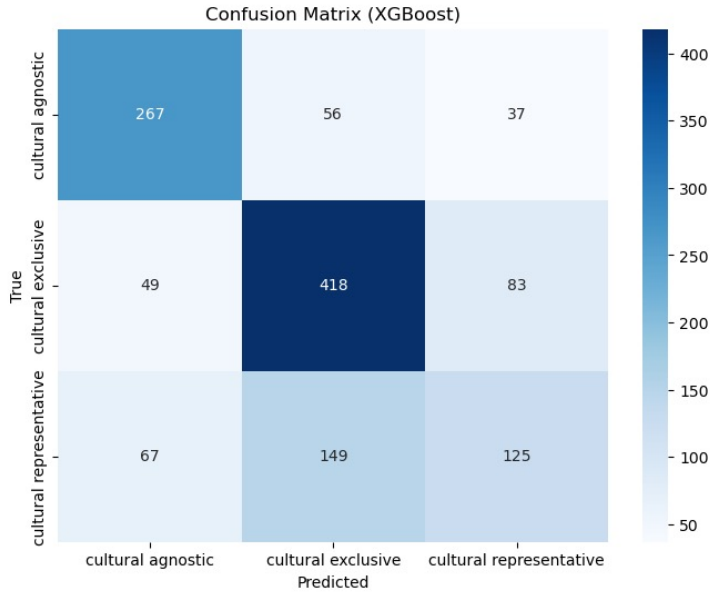
Model for the non-LLM task

The model chosen for this task was **XGBoost**

This model is based on decision trees and applies a different weight on each feature based on how many times it appears as a splitting condition

Class	Precision	Recall	F1 Score
CA	0.76	0.87	0.81
CE	0.63	0.78	0.69
CR	0.75	0.50	0.60

Overall Accuracy: 72%



The most confused class is the cultural representative one, as expected, since it is the one that shares the most amount properties with all the others

Homework 2: Ancient to modern Italian automatic translation

LLMs (qwen3, LLama3)

Two different prompting techniques used to generate the translations

Finetuning

Finetuning a machine translation model using a gold annotated dataset

Combined Approach

Usage of the highest scored LLM to correct the transaltions obtained via the finetuned model

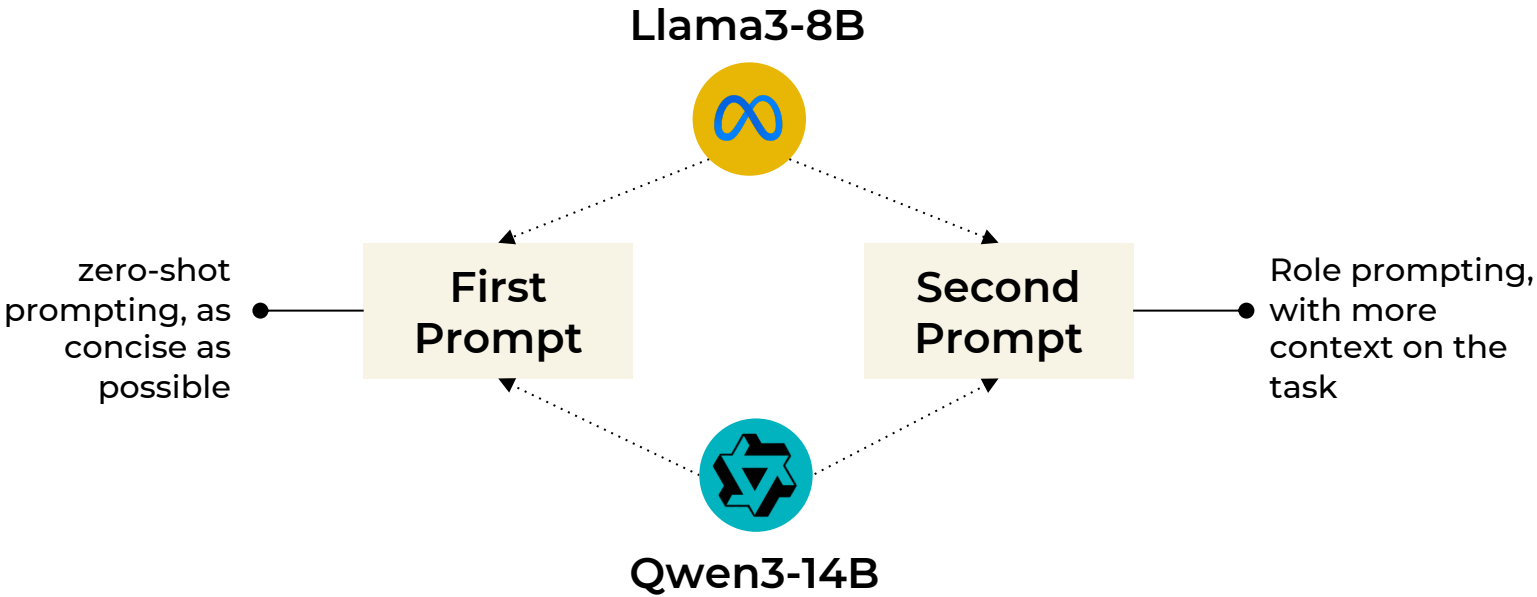
LLM as a judge

Grading the transaltions using Gemini 2.0 flash and Prometheus as judges

Human Correlation

Comparing the LLM's scores to the human's scores

LLM Translations



Original Sentence: *«Marco Cornelio ch'era de dieci compagni si riservò di parlare all'ultimo con studio»*

Llama translation: *«Marco Cornelio che era dei dieci compagni si riservò di parlare all'ultimo con studio»*

Qwen translation: *«Marco Cornelio che era uno dei dieci compagni si riservò di parlare per ultimo con attenzione»*

Finetuning

The finetuned model was a machine translation model **facebook/mbart-large-50-many-to-many-mmt**

A new gold annotated dataset was introduced

Sentence	Translation
Nel mezzo del cammin di nostra vita …	A metà del percorso della vita umana …
…	…

162 rows

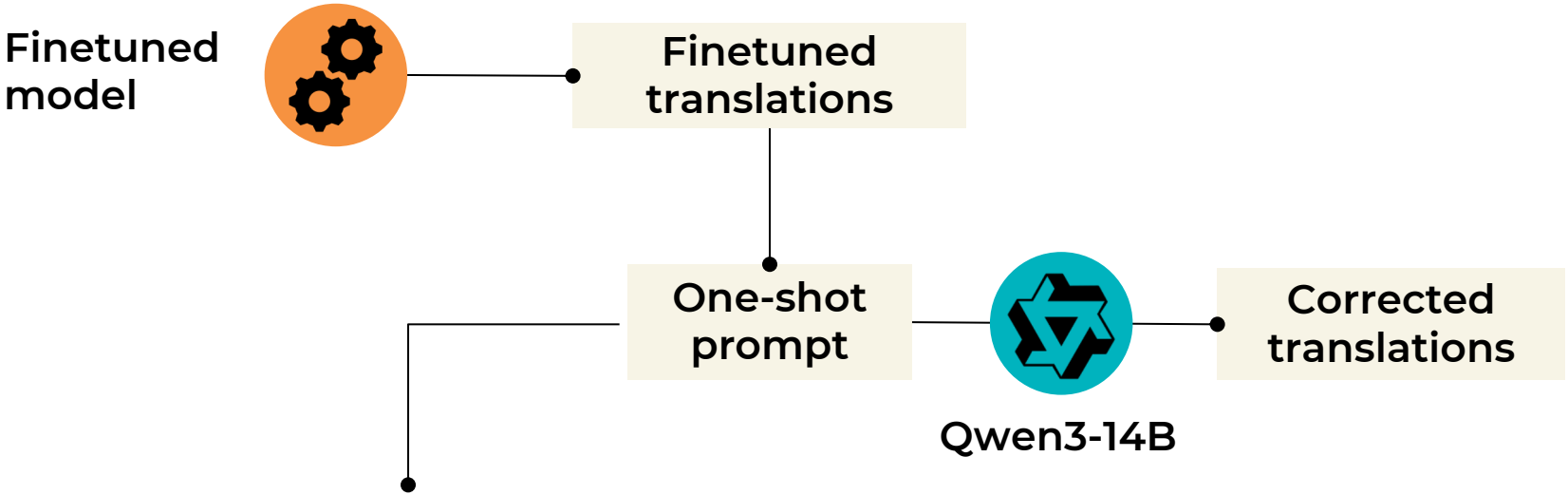
The dataset contains original non-modern Italian sentences and their official translations taken from:

- *«Decameron»*, Giovanni Boccaccio
- *«Divina Commedia, Inferno»*, Dante Alighieri
- *«Canzoniere»*, Francesco Petrarca
- *«Cantico delle Creature»*, San Francesco d'Assisi
- *«Donna Me Prega»*, Guido Cavalcanti
- *«Io voglio del ver la mia donna laudare»*, Guido Guinizzelli

Since the model used was a machine translation model the task was approached as a paraphrasing task

Epoch	Train Loss	Val Loss	Rouge-1	Rouge-2	Rouge-L	Rouge-L-sum
1	11.380	9.315	0.357	0.123	0.312	0.310
2	8.551	7.812	0.393	0.151	0.343	0.341
3	7.719	6.623	0.383	0.140	0.341	0.339
4	6.382	5.833	0.394	0.154	0.348	0.346
5	5.695	5.486	0.396	0.148	0.347	0.345

Combined Approach



«Revise the translation of the following sentence in old Italian maintaining as much as possible relevance to the original sentence's meaning.
Paraphrase the translation using modern italian terms, an example of acceptable correction is:

***Old Italian**: Marco Cornelio ch'era de' dieci compagni, studiosamente si riservò di parlare all'ultimo.*
***Translation**: Marco Cornelio che era dei dieci compagni, si riservò di parlare all'ultimo con studio.*
***Corrected Translation**: Marco Cornelio, che era tra i dieci compagni, scelse di parlare per ultimo intenzionalmente.»*

LLM as a Judge

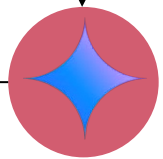
Grading System

Scoring system from 1 to 5 based mainly on syntactic and semantic errors

Prompt integration

Rubric integration

More lenient judge



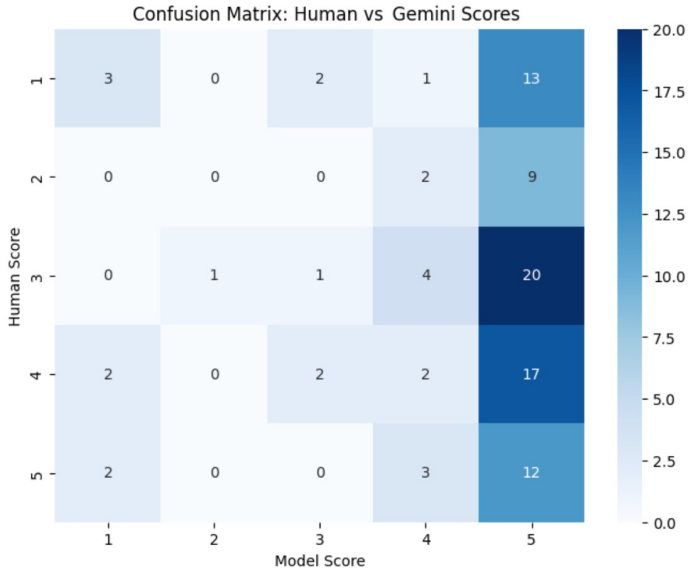
Gemini 2.0-flash

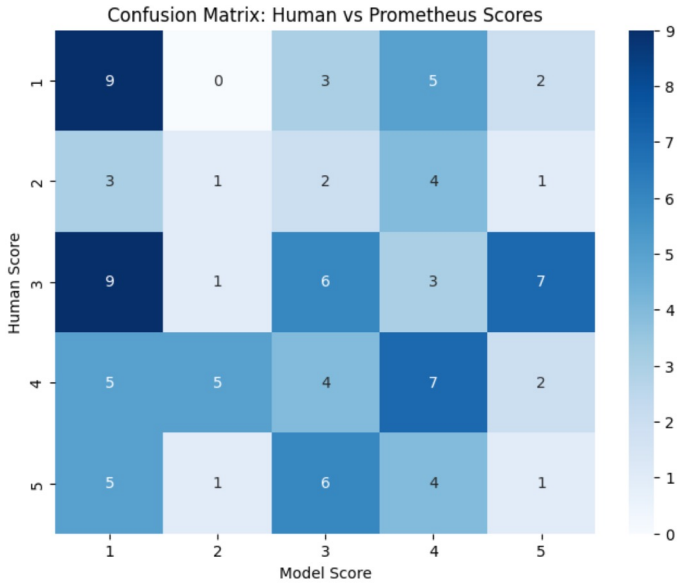
Prometheus-Eval

Harsher judge

Human Metrics Correlation

Considering now only the highest graded model: Qwen3-14B with the 1st prompt



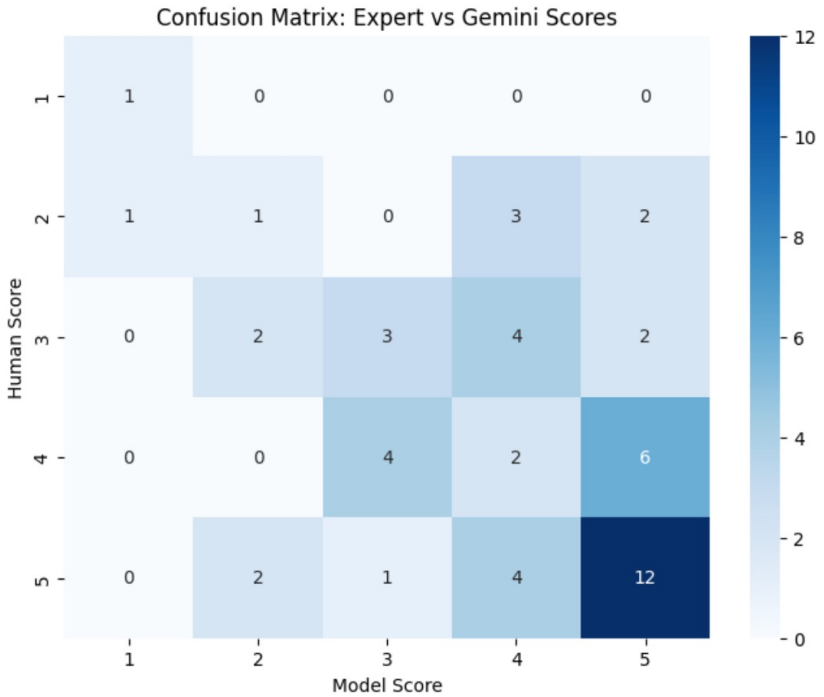


Model	Cohen's Kappa Coefficient
Gemini	0.02483
Prometheus	0.051399

Expert Metrics Correlation

Regarding our combined approach. we decided to ask an **expert in the field** to evaluate our translations

Recent Master's graduate in «Filologia Moderna LM-14» at La Sapienza



Cohen's Kappa Coefficient		
Expert vs LLM	Human vs LLM	Expert vs Human
0.396	0.504	0.444

Thank you for listening