Large Language Models For Text Classification: Case Study And Comprehensive Review

Arina Kostina - akosti02@ucy.ac.cy
Marios D. Dikaiakos - mdd@cs.ucy.ac.cy
Dimosthenis Stefanidis - dstefa02@ucy.ac.cy
George Pallis - gpallis@cs.ucy.ac.cy

University of Cyprus

Motivation and Research Questions

Unlocking the potential of Large Language Models (LLMs) in data classification represents a promising frontier in natural language processing.
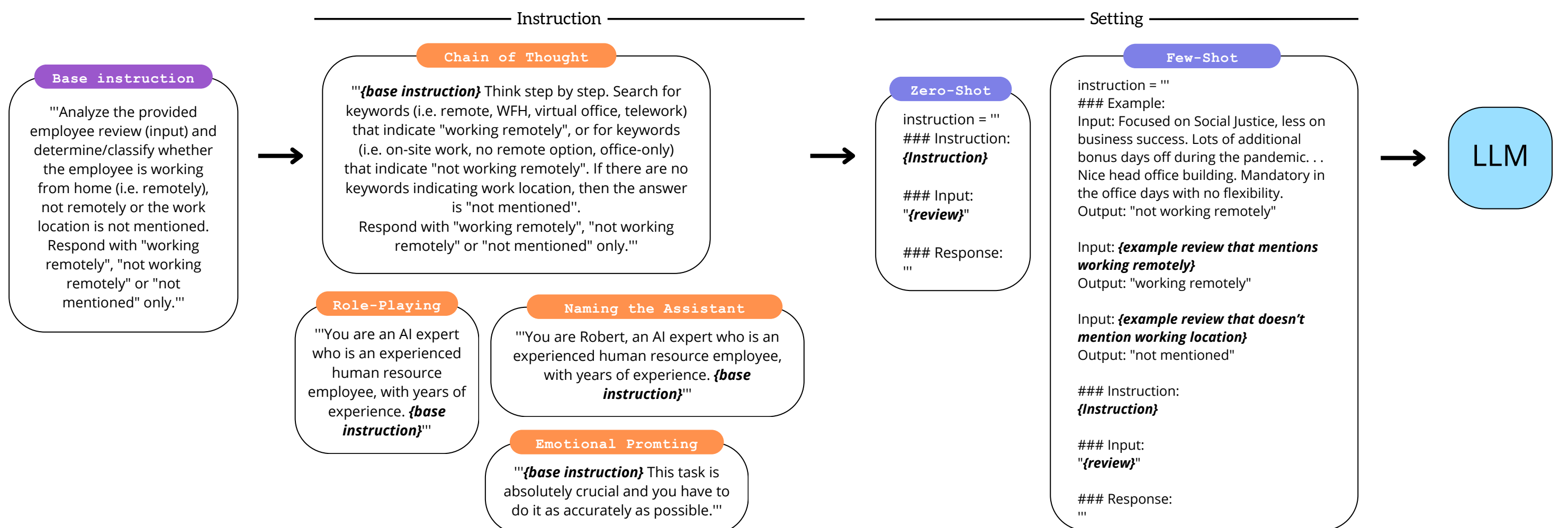
RQ1: Evaluate the capabilities of open-source quantized LLM models and compare their performance against traditional state-of-the-art method roBERTa in the task of data classification

RQ2: Explore how factors like model scale, base models, and prompting techniques, influence classification results

Prompting Techniques

Employee Review Example:
Great People, Great Culture. I've worked with a lot of people and have not worked with a more supportive/responsive remote team at any other past job. The work culture is also great. Lots of PTO that people actually use and a general respect for life outside of work.
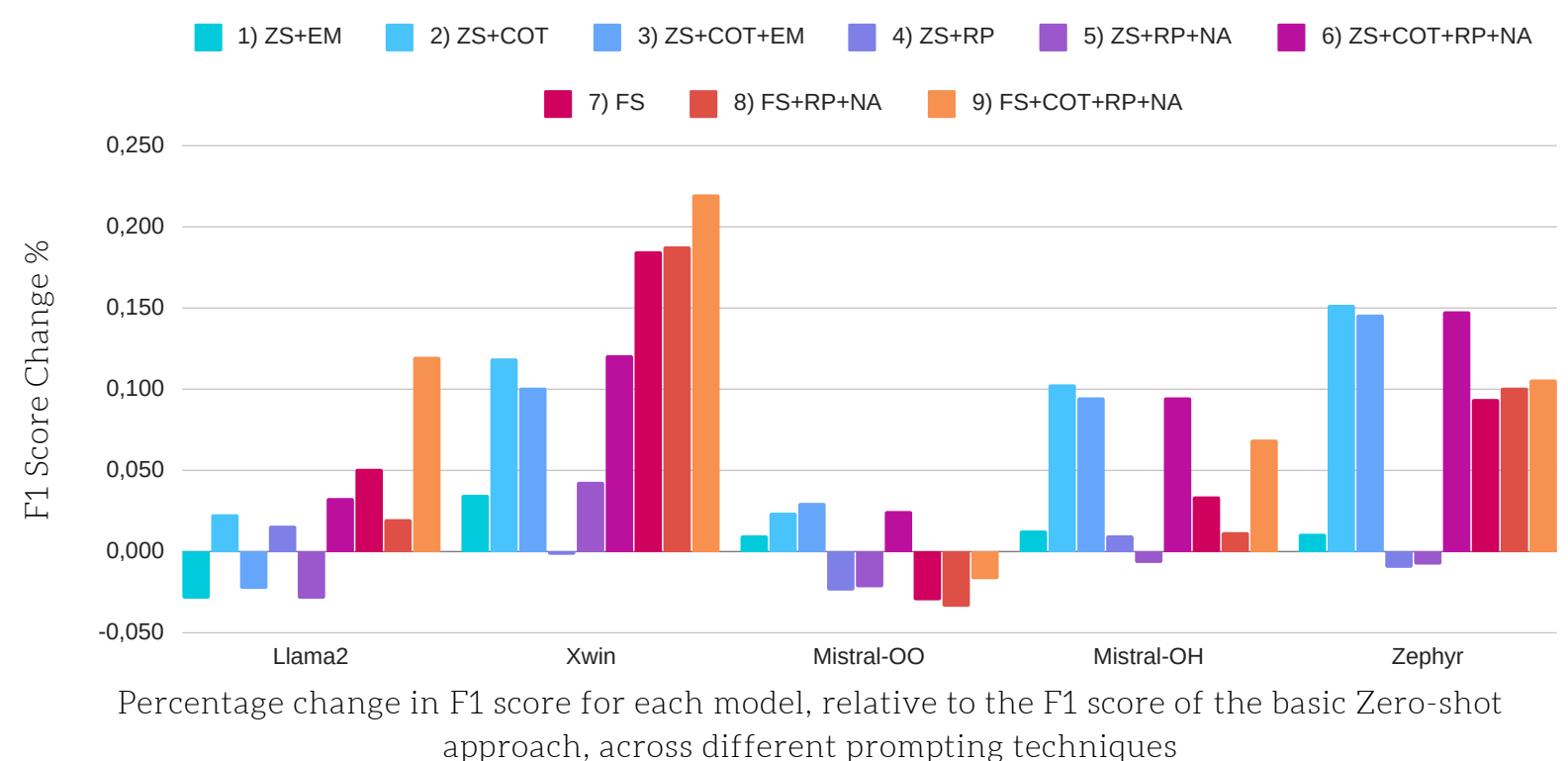
Experimental Setup

Classification Task: Classify large amount of employee company reviews based on their working location.

Data: Company reviews from the Glassdoor website, where current and former employees anonymously review companies and their management.

LLM models: Mistral-7B OpenOrca *(Mistral-OO)*, OpenHermes 2.5 Mistral-7B *(Mistral-OH)*, zephyr-7B-beta *(Zephyr)*, Nous-Hermes Llama2 13B *(Llama2)*, Xwin-MLewd 13B v0.2 *(Xwin)*

Testing Dataset and Categories: Manually annotated sample of 1000 reviews, with 37% reviews in "working remotely", 28% in "not working remotely", 35% in "not mentioned"

Instruction — Setting

Base instruction
'''Analyze the provided employee review (input) and determine/classify whether the employee is working from home (i.e. remotely), not remotely or the work location is not mentioned. Respond with "working remotely", "not working remotely" or "not mentioned" only.'''

Chain of Thought
'''*{base instruction}* Think step by step. Search for keywords (i.e. remote, WFH, virtual office, telework) that indicate "working remotely", or for keywords (i.e. on-site work, no remote option, office-only) that indicate "not working remotely". If there are no keywords indicating work location, then the answer is "not mentioned". Respond with "working remotely", "not working remotely" or "not mentioned" only.'''

Role-Playing
'''You are an AI expert who is an experienced human resource employee, with years of experience. *{base instruction}*'''

Naming the Assistant
'''You are Robert, an AI expert who is an experienced human resource employee, with years of experience. *{base instruction}*'''

Emotional Promting
'''*{base instruction}* This task is absolutely crucial and you have to do it as accurately as possible.'''

Zero-Shot
instruction = ''
Instruction:
{Instruction}

Input:
"*{review}*"

Response:
''

Few-Shot
instruction = ''
Example:
Input: Focused on Social Justice, less on business success. Lots of additional bonus days off during the pandemic. . . Nice head office building. Mandatory in the office days with no flexibility.
Output: "not working remotely"

Input: *{example review that mentions working remotely}*
Output: "working remotely"

Input: *{example review that doesn't mention working location}*
Output: "not mentioned"

Instruction:
{Instruction}

Input:
"*{review}*"

Response:
''

LLM

Results and Observations

RQ1:
- **roBERTa** achieved **85.5%** F1 score
- **Mistral-OO** achieved the highest performance with an F1 score of **86.4%, beating roBERTa**

RQ2:
- **By utilizing different prompting techniques** that try to stimulate reasoning, models can achieve as much as a **22.2% point increase.**
- **Larger LLMs (13B) perform better in the Few-Shot setting than in Zero-Shot setting** indicating that they can utilize the information provided in the examples more effectively
- **The Chain-of-Thought technique and the Few-Shot setting** are able to offer notable performance increase
- **Mistral-based models consistently showcase better performance** than the Llama2-based models
- **Mistral-OO** and **Mistral-OH** have different performances, with their **only difference lying in the training dataset**

1) ZS+EM 2) ZS+COT 3) ZS+COT+EM 4) ZS+RP 5) ZS+RP+NA 6) ZS+COT+RP+NA
7) FS 8) FS+RP+NA 9) FS+COT+RP+NA

Percentage change in F1 score for each model, relative to the F1 score of the basic Zero-shot approach, across different prompting techniques

		Llama2	Xwin	Mistral-OO	Mistral-OH	Zephyr	Roberta
0	Zero-shot	0,507	0,522	0,834	0,716	0,611	
1	Zero-shot + Emotional prompting	0,478	0,557	0,844	0,729	0,622	
2	Zero-shot + CoT	0,53	0,641	0,858	**0,819**	**0,763**	
3	Zero-shot + CoT + Emotional prompting	0,484	0,623	**0,864**	0,811	0,757	
4	Zero-shot + Role playing	0,523	0,52	0,81	0,726	0,601	
5	Zero-shot + Role playing + Naming the Assistant	0,478	0,565	0,812	0,709	0,603	**0,855**
6	Zero-shot + CoT + Role playing + Naming the Assistant	0,54	0,643	0,859	0,811	0,759	
7	Few-shot	0,558	0,707	0,804	0,75	0,705	
8	Few-shot + Role playing + Naming the Assistant	0,527	0,71	0,8	0,728	0,712	
9	Few-shot + CoT + Role playing + Naming the Assistant	**0,627**	**0,742**	0,817	0,785	0,717	

Heatmap of F1 Scores for Each Model Across Different Prompts