Will LLMs Replace the Encoder-Only Models in Temporal Relation Classification?

Gabriel Roccabruna, Massimo Rizzoli, Giuseppe Riccardi
Signals and Interactive Systems Lab, University of Trento, Italy

BACKGROUND

Temporal relations order events chronologically

- ***Encoder-only models*** still achieve SOTA performance in the TRC task
- We know little about the ***performance of LLMs***

Temporal Relation Classifcation

It **e_1: accused** the company of deliberately slashing oil revenues by overproducing oil and **e_2: driving** down prices, among other charges.

AFTER

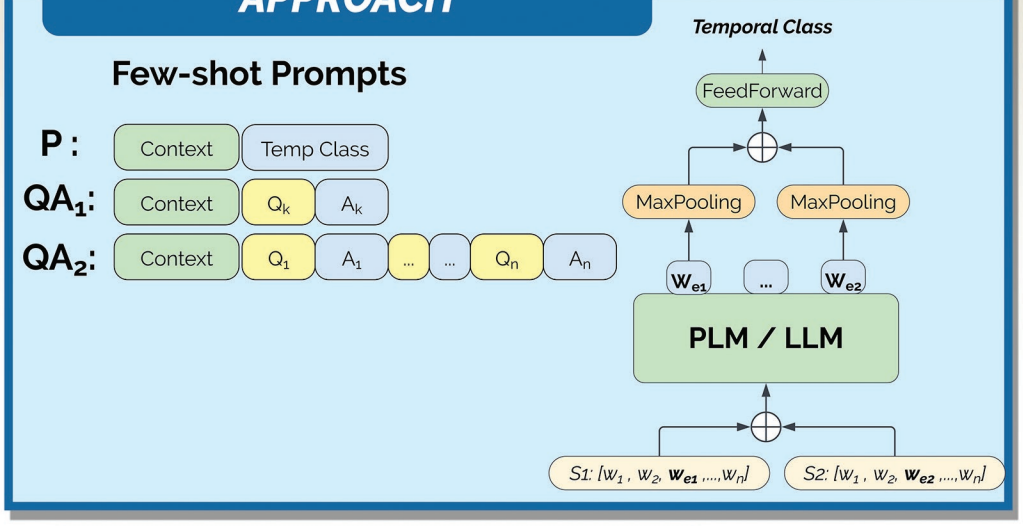
e_1: accused → **e_2: driving**

CONTRIBUTIONS

"Are LLMs as good as Encoder-only models at modelling Temporal Relations?"

1. Evaluating LLMs with ICL and fine-tuning
2. Investigating reasons behind the difference in performance between LLMs and Encoder models

APPROACH



RESULTS

Models	MATRES			TIMELINE			TB-Dense		
	P	*QA₁*	*QA₂*	*P*	*QA₁*	*QA₂*	*P*	*QA₁*	*QA₂*
Mistral 7B	30.0	14.8	52.9	28.7	8.1	39.9	5.0	0.4	0.0
Mixtral 8×7B	27.7	28.1	58.0	36.1	30.2	53.2	8.5	12.3	13.1
Llama2 7B	31.2	14.8	56.3	41.8	9.7	58.1	21.7	1.6	0.6
Llama2 13B	36.7	8.5	31.1	41.8	8.0	28.3	27.9	3.3	24.3
Llama2 70B	36.6	37.0	**65.3**	39.4	48.0	**62.5**	27.1	9.3	**31.4**
GPT-3	54.0	8.0	55.6	7.0	20.3	57.3	2.7	2.5	0.5
GPT-3.5	41.2	29.6	61.2	11.7	12.2	58.5	19.0	24.6	12
Llama2 7B$_{\text{Fine-tuned}}$	71.4	77.2	82.0	57.2	**76.9**	55.9	45.0	4.7	49.3
Llama2 13B$_{\text{Fine-tuned}}$	76.5	81.6	**84.3**	61.3	30.5	41.5	**55.4**	3.7	48.7
RoBERTa	87.6			87.9			83.1		

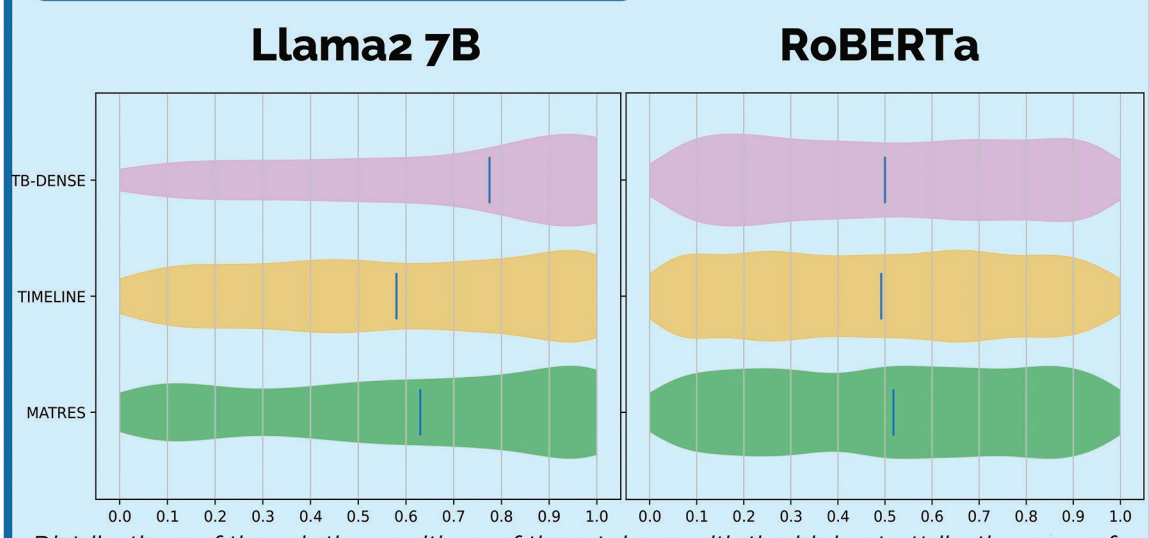
WORD EMB. ANALYSIS

Models	Frozen Encoder			Full Fine-Tuning		
	MATR.	TIMEL.	TB-D.	MATR.	TIMEL.	TB-D.
Llama2 7B	75.2	64.8	68.0	79.4	64.9	77.3
Llama2 13B	76.6	66.6	68.9	82.8	69.8	77.7
Llama2 70B	75.9	**69.1**	65.7	81.5	67.2	72.4
RoBERTa	**80.5**	65.7	**71.4**	**87.6**	**87.9**	**83.1**

Event representations computed using a ***frozen RoBERTa*** are ***more effective*** in the TRC task than those computed with LLMs.

Training the encoder is ***effective also for LLMs*** but RoBERTa still achieves the best performance.

XAI



Distributions of the relative positions of the 5 tokens with the highest attribution score for each sequence.

1) LLama2 7B tends to focus on the last tokens.
2) RoBERTa exploits the whole sequence more uniformly.

CONCLUSIONS

Encoder-only models **outperform** LLMs in the TRC task.

Our analyses suggest that this could be due to the **different pre-training** tasks.

Overall, a **more accurate** and **low-resource** demanding ***RoBERTa-based*** model should be preferred ***over an LLM*** in the TRC task.

SCAN ME!

EMNLP 2024

UNIVERSITY OF TRENTO

ICT DAYS

UNIVERSITÀ DI TRENTO
Dipartimento di Ingegneria e Scienza dell'Informazione

SiS LAB
Signals & Interactive Systems