LREC-COLING 2024

Analyzing Large Language Models' Capability in Location Prediction

***Zhaomin Xiao, *Yan Huang, and **Eduardo Blanco**

*Computer Science and Engineering, University of North Texas
**Computer Science, University of Arizona
zhaominxiao@my.unt.edu, yan.huang@unt.edu, eduardoblanco@arizona.edu

Motivation

- LLMs have gained substantial attraction since GPT-3.
  - ChatGPT, LLaMA 2, Gemini, etc.
- They perform pretty well in various NLP tasks.
- No previous effort on location prediction.

Contributions

Contributions

- We show that LLMs can perform the task of deciding whether users were located in the location they mentioned in tweets

Contributions

- We show that LLMs can perform the task of deciding whether users were located in the location they mentioned in tweets
- Our experiments show that
  - Instruction finetuning is not consistently beneficial
  - Providing exemplars is generally helpful
  - Considering the tweet published before and after the tweet mentioning the location is not always beneficial in the context of LLMs

Contributions

- We show that LLMs can perform the task of deciding whether users were located in the location they mentioned in tweets
- Our experiments show that
  - Instruction finetuning is not consistently beneficial
  - Providing exemplars is generally helpful
  - Considering the tweet published before and after the tweet mentioning the location is not always beneficial in the context of LLMs
- Ablation study shows that some modifications of instructions are essential to achieve the best results

Contributions

- We show that LLMs can perform the task of deciding whether users were located in the location they mentioned in tweets
- Our experiments show that
  - Instruction finetuning is not consistently beneficial
  - Providing exemplars is generally helpful
  - Considering the tweet published before and after the tweet mentioning the location is not always beneficial in the context of LLMs
- Ablation study shows that some modifications of instructions are essential to achieve the best results
- Analyses of the errors made by the best-performing model

Task & Dataset

Task & Dataset

Given a tweet mentioning a location, determine whether the author of the tweet was there when the tweet was published.

Task & Dataset

Given a tweet mentioning a location, determine whether the author of the tweet was there when the tweet was published.

`Yes`: The author of the tweet *was* at Dallas when the tweet was published.

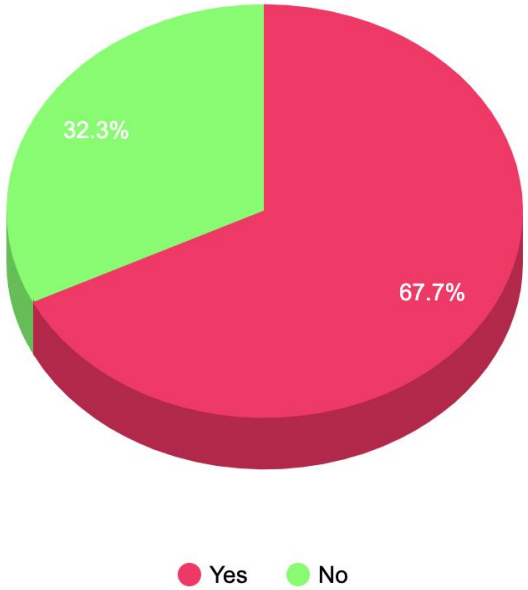
`No`: *I cannot determine* whether the author of the tweet was at Dallas when the tweet was published.

Task & Dataset

3,494 instances in total

each instance → 3 earlier tweets
1 target tweet (mentions the location)
3 later tweets

2/3 `Yes` & 1/3 `No`



Task & Dataset

- Broad population coverage
  - Large cities: 82.1% (e.g., Chicago and Miami)
  - Small cities: 17.9% (e.g., Reno and Toledo)
- Broad geographical coverage
  - Northeastern: 10.4% (e.g., Massachusetts)
  - Midwestern: 15.5% (e.g., Missouri)
  - Southern: 54.9% (e.g., Texas)
  - Western: 19.2% (e.g., California)

Experiments

- LLM selection
  - Encoder-decoder-based models
    - FLAN-T5 & FLAN-UL2
  - Decoder-based models
    - FLAN-Alpaca
  - Closed & Commercial models
    - ChatGPT

Experiments

- Prompt design
  - Instruction modifications
  - Add flags to indicate the ordering of each tweet when showing more than target tweets

Read the tweet and determine if the author of the tweet was located at <loc> when the tweet was published. The '#' in the hashtags and '@' in the mentions are removed. If the tweet is associated with advertisements or news reports, then you can be more confident in selecting yes.
<tweet_text>
1. yes, the author of the tweet was located at <loc> when the tweet was published. 2. no, I cannot determine if the author of the tweet was located at <loc> when the tweet was published.

Table 2: Our prompt for location prediction. <loc> and <tweet_text> are the mentioned location and the text of the tweet, respectively.

Experiments

- Instruction finetuning
  - LoRA
    - Freeze LLMs’ weights
    - Introduce trainable matrices for each layer
    - Greatly reduce the number of trainable parameters
  - ChatGPT’s API

Results

Finding #1: Instruction finetuning is not consistently beneficial

Model		0-shot	1-shot	5-shot	10-shot
Majority baseline		0.55			
Without instruction finetuning	ChatGPT	0.48	0.57	0.57	0.57
	FLAN-T5	0.38	0.44	0.48	0.50
	FLAN-Alpaca	0.17	0.40	0.47	0.48
	FLAN-UL2	0.59	0.60	**0.62**	0.61
With instruction finetuning	ChatGPT	0.58	0.59	0.61	0.60
	FLAN-T5	0.55	0.55	0.59	0.59
	FLAN-Alpaca	0.27	0.33	0.50	0.55
	FLAN-UL2	0.58	0.57	0.54	0.53

Results

Finding #1: Instruction finetuning is not consistently beneficial

Instruction finetuning works!

Model		0-shot	1-shot	5-shot	10-shot
Majority baseline		0.55			
Without instruction finetuning	ChatGPT	0.48	0.57	0.57	0.57
	FLAN-T5	0.38	0.44	0.48	0.50
	FLAN-Alpaca	0.17	0.40	0.47	0.48
	FLAN-UL2	0.59	0.60	**0.62**	0.61
With instruction finetuning	ChatGPT	0.58	0.59	0.61	0.60
	FLAN-T5	0.55	0.55	0.59	0.59
	FLAN-Alpaca	0.27	0.33	0.50	0.55
	FLAN-UL2	0.58	0.57	0.54	0.53

Results

Finding #1: Instruction finetuning is not consistently beneficial

But it does not always work!

Model		0-shot	1-shot	5-shot	10-shot
Majority baseline		0.55			
Without instruction finetuning	ChatGPT	0.48	0.57	0.57	0.57
	FLAN-T5	0.38	0.44	0.48	0.50
	FLAN-Alpaca	0.17	0.40	0.47	0.48
	FLAN-UL2	0.59	0.60	**0.62**	0.61
With instruction finetuning	ChatGPT	0.58	0.59	0.61	0.60
	FLAN-T5	0.55	0.55	0.59	0.59
	FLAN-Alpaca	0.27	0.33	0.50	0.55
	FLAN-UL2	0.58	0.57	0.54	0.53

Results

Finding #2: Providing exemplars mostly helps

Model		0-shot	1-shot	5-shot	10-shot
Majority baseline		0.55			
Without instruction finetuning	ChatGPT	0.48	0.57	0.57	0.57
	FLAN-T5	0.38	0.44	0.48	0.50
	FLAN-Alpaca	0.17	0.40	0.47	0.48
	FLAN-UL2	0.59	0.60	**0.62**	0.61
With instruction finetuning	ChatGPT	0.58	0.59	0.61	0.60
	FLAN-T5	0.55	0.55	0.59	0.59
	FLAN-Alpaca	0.27	0.33	0.50	0.55
	FLAN-UL2	0.58	0.57	0.54	0.53

Results

Finding #2: Providing exemplars mostly helps

more exemplars -> better results

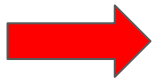
Model		0-shot	1-shot	5-shot	10-shot
Majority baseline					
Without instruction finetuning	ChatGPT	0.48	0.57	0.57	0.57
	FLAN-T5	0.38	0.44	0.48	0.50
	FLAN-Alpaca	0.17	0.40	0.47	0.48
	FLAN-UL2	0.59	0.60	**0.62**	0.61
With instruction finetuning	ChatGPT	0.58	0.59	0.61	0.60
	FLAN-T5	0.55	0.55	0.59	0.59
	FLAN-Alpaca	0.27	0.33	0.50	0.55
	FLAN-UL2	0.58	0.57	0.54	0.53

Results

Finding #2: Providing exemplars mostly helps

Model		0-shot	1-shot	5-shot	10-shot
Majority baseline		0.55			
Without instruction finetuning	ChatGPT	0.48	0.57	0.57	0.57
	FLAN-T5	0.38	0.44	0.48	0.50
	FLAN-Alpaca	0.17	0.40	0.47	0.48
	FLAN-UL2	0.59	0.60	**0.62**	0.61
With instruction finetuning	ChatGPT	0.58	0.59	0.61	0.60
	FLAN-T5	0.55	0.55	0.59	0.59
	FLAN-Alpaca	0.27	0.33	0.50	0.55
	FLAN-UL2	0.58	0.57	0.54	0.53

Too many exemplars are not good



Results

Finding #3: Context is not always helpful

Model	*Target*	*Earlier+Target*	*Target+Later*	*All*
ChatGPT	0.57	0.59	**0.61**	**0.61**
FLAN-T5	**0.48**	0.41	0.41	0.42
FLAN-Alpaca	**0.47**	0.39	0.40	0.39
FLAN-UL2	**0.62**	0.58	0.59	0.59

Results

Finding #3: Context is not always helpful

Context is selectively helpful

Model	*Target*	*Earlier+Target*	*Target+Later*	*All*
ChatGPT	0.57	0.59	**0.61**	**0.61**
FLAN-T5	**0.48**	0.41	0.41	0.42
FLAN-Alpaca	**0.47**	0.39	0.40	0.39
FLAN-UL2	**0.62**	0.58	0.59	0.59

Results

Finding #3: Context is not always helpful

Model	*Target*	*Earlier+Target*	*Target+Later*	*All*
ChatGPT	0.57	0.59	**0.61**	**0.61**
FLAN-T5	**0.48**	0.41	0.41	0.42
FLAN-Alpaca	**0.47**	0.39	0.40	0.39
FLAN-UL2	**0.62**	0.58	0.59	0.59

Target tweets are enough, more context degrade model performance

Results

Finding #3: Context is not always helpful

Earlier tweet: I'm in Denver for spring break. Saw this walking to lunch. It's at a restaurant called Pride and Swagger. #SaLuna #SamandLunaForever
Target tweet: @GuyFieri I'm in Denver for my spring break. I went to Steuben's for lunch today. I had the blt, and it was incredible!
Later tweet: @MLB_PR @MLB @AtlanticLg I don't get why they don't try this in their affiliated leagues.

Table 4: Example showing that taking into account context tweets is not beneficial.

Results

Finding #3: Context is not always helpful

Target tweet: @GuyFieri I'm in Denver for my spring break. I went to Steuben's for lunch today. I had the blt, and it was incredible!

Table 4: Example showing that taking into account context tweets is not beneficial.

Results

Finding #3: Context is not always helpful

Target tweet: @GuyFieri I'm in Denver for my spring break. I went to Steuben's for lunch today. I had the blt, and it was incredible!

Table 4: Example showing that taking into account context tweets is not beneficial.

The author was at Denver

Results

Finding #3: Context is not always helpful

No need to look at the context

Earlier tweet: I'm in Denver for spring break. Saw this walking to lunch. It's at a restaurant called Pride and Swagger. #SaLuna #SamandLunaForever
Target tweet: @GuyFieri I'm in Denver for my spring break. I went to Steuben's for lunch today. I had the blt, and it was incredible!
Later tweet: @MLB_PR @MLB @AtlanticLg I don't get why they don't try this in their affiliated leagues.

Table 4: Example showing that taking into account context tweets is not beneficial.

The author was at Denver

Ablation Study

Read the tweet and determine if the author of the tweet was located at <loc> when the tweet was published. The '#' in the hashtags and '@' in the mentions are removed. If the tweet is associated with advertisements or news reports, then you can be more confident in selecting yes.

<tweet_text>

1. yes, the author of the tweet was located at <loc> when the tweet was published.
2. no, I cannot determine if the author of the tweet was located at <loc> when the tweet was published.

Table 2: Our prompt for location prediction. <loc> and <tweet_text> are the mentioned location and the text of the tweet, respectively.

FLAN-UL2 w/ both strategies	0.62
FLAN-UL2 w/o *preprocess*	0.60
FLAN-UL2 w/o *enhance*	0.60

preprocess: Remove "@" and "#" from tweets and provide corresponding messages in the instruction.

enhance: Add the clue to enhance the models' confidence (i.e, if the tweet is associated with ... in selecting `yes`).

Ablation Study

Read the tweet and determine if the author of the tweet was located at <loc> when the tweet was published. The '#' in the hashtags and '@' in the mentions are removed. If the tweet is associated with advertisements or news reports, then you can be more confident in selecting yes.

<tweet_text>

1. yes, the author of the tweet was located at <loc> when the tweet was published.
2. no, I cannot determine if the author of the tweet was located at <loc> when the tweet was published.

Table 2: Our prompt for location prediction. <loc> and <tweet_text> are the mentioned location and the text of the tweet, respectively.

FLAN-UL2 w/ both strategies	0.62
FLAN-UL2 w/o *preprocess*	0.60
FLAN-UL2 w/o *enhance*	0.60

preprocess: Remove "@" and "#" from tweets and provide corresponding messages in the instruction.

enhance: Add the clue to enhance the models' confidence (i.e, if the tweet is associated with ... in selecting `yes`).

Both strategies for instruction modification are essential

Qualitative Analysis

Error Type	Example
Ads/News content (48%)	The Ashland University Band performed over spring break among dinosaurs and elephants at the Field Museum during their Chicago tours! Let's show our hospitality as the host. Come and join us!!! #Tour #LocalBusiness Mentioned location: *Chicago*, Ground truth: `Yes`, Prediction: `No`
Irrelevant discussion (23%)	y'all coming back from Miami and Mexico after thottin and boppin all spring break with this rona outbreak Mentioned location: *Miami*, Ground truth: `No`, Prediction: `Yes`
Short text (14%)	Happy Thanksgiving to my #HeatNation @Shesk305 @Bballilluminous @dionwebster10 @HeatLoco @miaheatbeat @MiamiHEAT #Thanksgiving Mentioned location: *Miami*, Ground truth: `Yes`, Prediction: `No`

Table 6: Most common errors made by FLAN-UL2 using target tweets.

Conclusions

- LLMs can tackle this challenging task
  - Outperform the majority baseline
- We need to play with them
  - Tweet preprocessing
  - Instruction modification
  - Show/hide context
  - ...
- Other LLMs & instruction finetuning strategies?

Thank you!
Any questions?