Analyzing Large Language Models' Capability in Location Prediction

Zhaomin Xiao[1], Yan Huang[1], Eduardo Blanco[2]
[1]University of North Texas, [2]University of Arizona
zhaominxiao@my.unt.edu, yan.huang@unt.edu, eduardoblanco@arizona.edu

Abstract

In this paper, we investigate and evaluate large language models' capability in location prediction. We present experimental results with four models—FLAN-T5, FLAN-UL2, FLAN-Alpaca, and ChatGPT—in various instruction finetuning and exemplar settings. We analyze whether taking into account the context—tweets published before and after the tweet mentioning a location—is beneficial. Additionally, we conduct an ablation study to explore whether instruction modification is beneficial. Lastly, our qualitative analysis sheds light on the errors made by the best-performing model.

Keywords: Location Prediction, Large Language Models

1. Introduction

X[1] (formally Twitter) is a social media platform where users publish short messages called tweets. According to a recent report, it ranks as the 14th most widely used social network globally, with 237.8 million daily active users as of 2023 (Shewale, 2024). X users can share their locations when posting tweets, although this feature is disabled by default. Due to the sparsity of geotagged tweets, there is a growing body of research trying to determine people's location based on their posted tweets (Doudagiri et al., 2018; Xiao and Blanco, 2022; Lamsal et al., 2022).

Large language models (LLMs) have made huge progress and gained popularity since 2022 (Chowdhery et al., 2022; Touvron et al., 2023a; Le et al., 2022), but remain underutilized in the field of location prediction. Because LLMs are not trained to predict locations, it is not clear if LLMs treat locations mentioned in a tweet simply as text or could understand the context and use it to decide if a user is physically located there. This paper seeks to bridge this gap by utilizing LLMs to tackle the problem of location prediction with tweets.

The work in this paper plays an important role in many applications, such as public health and epidemiology (Delmelle et al., 2022), emergency responses, and urban planning (Casali et al., 2022). For example, the location information available on social media platforms can be utilized to offer personalized recommendations in the hospitality and tourism industry (Mirzaalian and Halpenny, 2019) and to prevent crimes (Monika and Bhat, 2022).

The main contributions of this paper are (a) we show that LLMs can perform the new task of deciding whether a user is located in the location mentioned in tweets; (b) we demonstrate that (b.1) instruction finetuning is not consistently beneficial in location prediction, (b.2) providing exemplars is generally helpful, and (b.3) taking into account the tweets published before and after the tweet mentioning the location is not always beneficial in the context of LLMs; (c) we conduct an ablation study to show that two strategies of instruction modification, tweet preprocessing (i.e., remove special characters from tweet text) and confidence enhancement (i.e., add the sentence in the instruction to force LLMs to be more confident), are beneficial; and (d) we perform a qualitative analysis to provide insights into the errors made by the best-performing LLM.

[1]http://twitter.com

2. Related Works

Most previous works on noisy user-generated content and spatial information focus on: a) named entity recognition (Shen et al., 2018; Ushio et al., 2022) and disambiguation (Inkpen et al., 2017; Hu et al., 2023), and b) location prediction (Li and Lim, 2022). The former identifies, among others, location named entities and links them to a knowledge base without specifying who is there. The latter, where our work falls into, aims to assign a location to the user.

Previous works targeting location prediction using X data are generally divided into two categories: a) home location prediction (Simanjuntak et al., 2022; Mostafa et al., 2022), where the objective is to predict long-term residential addresses of X users, and b) real-time location prediction (Lutsai and Lampert, 2023; Julie et al., 2023), whose goal is to predict the location where a tweet is posted in real-time as it gets published.

LLMs have excelled in various NLP tasks such as sentiment analysis (Bang et al., 2023), natural language inference (Lee et al., 2023), machine translation (Lyu et al., 2023), and question answering (Bai

LREC-COLING 2024, pages 951–958

et al., 2023). More related to our work, Wang et al. (2023) use LLMs with prompts designed to incorporate mobility data, such as historical stays and contextual stays, to predict human movements. Our research differs in two aspects. Firstly, we aim to determine whether the author of the tweet was located at the mentioned location. Secondly, we do not rely on user's past precise locations. Instead, we solely rely on the tweet text. To the best of our knowledge, we are the first to use LLMs and tweet text for location prediction.

3. Definitions and Background

3.1. Problem Definition

Given a pair of (t, l), where t is a tweet, and l is a location mentioned in t, the objective is to predict the spatial relationship between the author of t and l. Specifically, we assign `yes` to (t, l) if the author of t was located at l when t was published and `no` if it is not possible to determine whether the author of t was located at l when t was published.

3.2. Dataset Description

To analyze the capability of LLMs in location prediction, we conduct experiments on an existing corpus (Xiao et al., 2023) with spatial annotations. This dataset consists of 3,494 instances, with each instance containing seven tweets that were published in chronological order. We denote the first three tweets, the middle tweet, and the last three tweets as *earlier tweets*, *target tweet*, and *later tweets*, respectively. We denote *context tweets* as the combination of *earlier tweets* and *later tweets*. A city from a predefined set is mentioned in the *target tweet*, but may or may not be mentioned in *earlier tweets* and *later tweets*. Most instances are annotated `yes` (67.7%), indicating that the author of the tweets was located at the city mentioned in the *target tweet* when it was published. The remaining one-third are annotated `no` (32.3%), indicating that it is not possible to determine whether the author of tweets was located at the city mentioned in the *target tweet* when the *target tweet* was published. The annotations are collected via Amazon Mechanical Turk and then filtered by MACE (Hovy et al., 2013), a tool designed to identify unreliable annotators and remove their annotations.

This dataset has a broad coverage, with a diverse range of cities in terms of population and geographical locations. Among the 94 cities mentioned in the 3,494 target tweets, 82.1% are large cities (e.g., Chicago and Miami), while 17.9% are smaller cities (e.g., Reno and Toledo).[2] This distribution is consistent with our intuition, as most cities discussed in tweets are large cities. Additionally, this dataset not only considers population diversity but also spans cities located in various regions of the United States. Specifically, 10.4% of target tweets mention 7 cities in the northeastern states (e.g., Massachusetts), 15.5% mention 18 cities in the Midwestern states (e.g., Missouri), 54.9% mention 33 cities in the southern states (e.g., Texas), and 19.2% mention 36 cities in the western states (e.g., California). Consequently, diverse geographical and cultural contexts, as reflected in the language used in tweets, are incorporated to ensure geographical generalizability.

3.3. LLM Selection

To analyze LLMs' capability in location prediction, we choose publicly available LLMs from two main categories based on their architectures. The first category comprises encoder-decoder-based models, which are built on the vanilla Transformer model (Vaswani et al., 2017), consisting of two stacks of Transformer blocks as the encoder and decoder, respectively. We select T5 (Raffel et al., 2020a) and UL2 (Tay et al., 2023a) in this category. The second category consists of decoder-based models, leveraging unidirectional attention masks to ensure each input token attends only to past tokens. In this category, we choose to work with Alpaca (Rishi et al., 2021), a LLaMA-based model. Previous work shows that instruction finetuning can boost LLMs' performance in downstream tasks (Chung et al., 2022). Hence, we work with enhanced versions of these models, specifically FLAN-T5, FLAN-UL2 (Tay, 2023), and FLAN-Alpaca (Chia et al., 2023). All models have undergone instruction finetuning on the FLAN Collection (Longpre et al., 2023), which consists of 473 datasets used in 1,836 NLP tasks, where each task is manually rephrased as instructions for instruction finetuning. Besides open-sourced LLMs, we also include ChatGPT[3] in our evaluation to assess how a closed (and commercial) LLM performs in location prediction.

4. Experiments and Analyses

4.1. Experimental Setup

Prompt design is important to elicit LLMs' ability to understand language (Liu et al., 2023). Table 2 presents our prompt for location prediction, including the instruction, tweet text, and options, which are shown in order. Note that we remove "#" and

[2] Note that small and large cities refer to cities with a population of less than 300,000 and equal to or larger than 300,000, respectively.

[3] https://chat.openai.com/

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

Table 1: Weighted average F1 scores of various LLMs, obtained with different numbers of exemplars and finetuning settings.

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.

"@" from the tweet's content and provide the corresponding message in the instruction as we find out that LLMs often struggle to understand hashtags and mentions unless we remove those characters. We also add the clue related to the ads and news (i.e., "If the tweet is associated with advertisements or news reports, then you can be more confident in selecting yes.") to enhance models' confidence because in our preliminary experiments, we observe that LLMs tend to be too conservative in predicting `yes` even when the tweet's content is related to local ads or news.

We create stratified train and test splits (70% and 30%) and use LoRA (Hu et al., 2022) to perform instruction finetuning for all LLMs except ChatGPT.[4] LoRA is a technique that freezes LLMs' weights while introducing trainable rank decomposition matrices into each layer of LLMs, greatly reducing the number of trainable parameters. The attention dimension of LoRA and dropout probability for LoRA layers are set as 4 and 0.1, respectively. We use AdamW optimizer (Loshchilov and Hutter, 2019) for instruction finetuning, with initial learning rate, β_1, β_2 as 1e-05, 0.9, 0.999, respectively, categorical cross entropy as loss function, and batch size 4. We conduct all experiments using 4 NVIDIA A100-80GB GPUs, which take a total of 4 days.

4.2. Instruction Finetuning is not Consistently Beneficial

Table 1 presents the weighted average F1 scores for different numbers of exemplars and instruction finetuning settings. We observe that instruction finetuning consistently enhances ChatGPT and FLAN-T5, although the improvement brought by instruction finetuning becomes smaller as the number of provided exemplars increases. Taking ChatGPT as an example, it gains 0.1 increase in F1 in the 0-shot setting, while it only gains 0.03 in the 10-shot setting. For FLAN-T5, instruction finetuning enables it to gain 0.17 F1 without being provided any exemplar, while gaining only around 0.1 F1 when being provided 1/5/10 exemplars. FLAN-Alpaca also benefits from instruction finetuning across most settings, except in the 1-shot setting. The only "outlier" is FLAN-UL2, in which instruction finetuning even leads to a decline in performance, regardless of the number of exemplars provided. This discrepancy shows that the effect of instruction finetuning is tailor to specific LLMs, as some LLMs experience substantial improvements while others do not.

4.3. Providing Exemplars Helps in Most Cases

We observe that providing more exemplars can improve model performance in most cases. Specifically, prior to instruction finetuning, all LLMs show improvements with more exemplars, as F1 score increases from 0.48 to 0.57 for ChatGPT, from 0.38 to 0.50 for FLAN-T5, from 0.17 to 0.48 for FLAN-Alpaca, and from 0.59 to 0.62 for FLAN-UL2. We also find out that the degree of the improvement decreases. In fact, negative effects emerge when providing more than 5 exemplars. In contrast, after instruction finetuning, most LLMs, with the excep-

[4]We finetune ChatGPT using its official API. The hyperparameters of finetuning are not accessible to us.

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

Table 3: Weighted average F1 scores of various LLMs under the 5-shot setting, obtained by providing different context tweets.

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.

tion of FLAN-Alpaca, cannot gain much from being provided exemplars. For instance, the F1 score increases from 0.58 to 0.61 for ChatGPT, and from 0.55 to 0.59 for FLAN-T5. For FLAN-Alpaca, its F1 score sees a significant increase with the inclusion of more exemplars, going from 0.27 to 0.55. Regarding FLAN-UL2, we find out that with instruction finetuning, its performance gets worse as more exemplars are provided.

4.4. Context is not Always Beneficial with LLMs

To investigate whether considering context tweets is helpful in location prediction in the context of LLMs, we conduct experiments with LLMs using different context tweets. Specifically, we consider four settings: a) only use target tweets, b) use earlier and target tweets, c) use target and later tweets, and d) use all tweets. We also modify the instructions and tweets so that LLMs understand the temporal relationship among these tweets. Specifically, we a) explicitly state in the instructions that tweets are published chronologically, and b) add indicators (e.g., Tweet 1, Tweet 2, etc.) before each tweet to indicate its order.

Table 3 shows the results obtained by various LLMs with 5 exemplars using different context tweets, as the best result in Table 1 is achieved in the 5-shot setting. We find out that incorporating context tweets is not always beneficial. More specifically, including later tweets alongside target tweets boosts ChatGPT's F1 score from 0.57 to 0.61, although using additional earlier tweets does not lead to further improvements. On the contrary, the other LLMs (i.e., FLAN-T5, FLAN-Alpaca, and

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

Table 5: Weighted average F1 scores of different strategies for instruction modifications. Note that all results are obtained by FLAN-UL2 using only target tweets in the 5-shot setting, as this setting yields the best results in Table 1 and Table 3.

FLAN-UL2) cannot benefit from any type of context tweets, indicating the variations in the models' ability to leverage context tweets.

Table 4 provides an example illustrating why considering context tweets is not beneficial. The target tweet is in the middle, while one tweet selected from the earlier and later tweets is shown at the top and bottom, respectively. We only show the most informative tweets in earlier and later tweets, as the others do not contain much information that can be leveraged to determine people's location. The spatiotemporal information contained within the target tweet ("I'm in *Denver* ... I went to ... today") is sufficient for the model to predict `yes`, as the context tweets either do not contain spatial information ("I don't get ...") or are partially duplicated to the target tweet ("I'm in Denver for my spring break").

4.5. Ablation Study

To explore whether instruction modification is beneficial, we conduct an ablation study. Figure 5 shows the results of the experiments with different strategies of instruction modification. We denote *preprocess* and *enhance* as the strategies of tweet preprocessing (i.e., remove "@" and "#" from tweets and provide corresponding message in the instruction) and confidence enhancement (i.e., add the clue—"If the tweet is associated with ... in selecting yes." in the instruction), respectively. We observe that the F1 score of FLAN-UL2 decreases from 0.62 to 0.60 when either tweet preprocessing or confidence enhancement is absent. This shows that both strategies of instruction modification are essential to obtain the best result.

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.

5. Qualitative Analysis

To better understand the errors made by the best-performing model, we conduct a qualitative analysis. We randomly select 100 errors made by FLAN-UL2 with 5 exemplars, using only target tweets, as the best results are yielded with this setting.

The most common error (48%) occurs when the tweet is associated with news or advertisements. The tweet at the top of Table 6 exemplifies this scenario. Sharing news about the local concert performance in Chicago ("Chicago tours", "our hospitality as the host", and "#LocalBusiness") provides a strong clue showing that the author of the tweet was in *Chicago* when the tweet was published. However, the model is too conservative to interpret this event as a local event and assign `yes` to this tweet.

The second most common error (23%) occurs when the discussion in the tweet is irrelevant to the tweet's author. The middle tweet in Table 6 illustrates this situation. Although mentioning *Miami*, the author is referring to other people ("y'all"). The model incorrectly identifies the tweet's subject, assuming that the author is still in Miami ("coming back from Miami") and predicts `yes`.

The third most common error (14%) takes place when the tweet text is too short. The "pure text" (text without hashtags and mentions, i.e., "Happy Thanksgiving to my") is too short to contain any spatial information. Additionally, the LLM also struggles to understand mentions, even without the character "@" (e.g., "Bballilluminous" and "dionwebster10"). Hence, it predicts `yes` as it mistakenly assumes that the author of the tweet was in *Miami* since Miami-related terms are frequently mentioned ("miaheatbeat" and "MiamiHEAT").

6. Conclusion

We have conducted extensive experiments to analyze LLMs' capability in location prediction. Our experimental results show that although providing exemplars generally help, instruction finetuning is not consistently beneficial, and the best results are achieved by FLAN-UL2 in the 5-shot setting without instruction finetuning. Our results and examples also show that in the context of LLMs, considering context tweets is not beneficial in most cases. The ablation study shows that both strategies for instruction modification, tweet preprocessing and confidence enhancement, are needed to obtain the best results. Lastly, our qualitative analysis provides insights into the errors made by the best-performing model.

7. Ethical Considerations

Location prediction has the potential for misuse, such as malicious tracking and surveillance. Applications that collect location data could sell that data to third parties, which can have serious implications for privacy. However, location prediction also provides multiple advantages, including enhanced user experiences, efficient marketing, resource management, safety improvements, and navigation enhancements.

We do not aim at tracking or surveillance. Instead, we focus on analyzing the LLMs' capability in location prediction. To address potential concerns, we have implemented the following safeguards:

1. The corpus we use contains only seven tweets per user published, making tracking and surveillance impossible. Additionally, neither user information nor any metadata is included in it.

2. Our experiments and analyses only take into account the tweet text. Additionally, in terms of LLMs' outputs, generating malicious content is also impossible because our carefully designed prompt constrains LLMs' output spaces (i.e., only select from the provided options).

8. Bibliographical References

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. 2023. Benchmarking foundation models with language-model-as-an-examiner. *ArXiv*, abs/2306.04181.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *ArXiv*, abs/2302.04023.

Ylenia Casali, Nazli Yonca Aydin, and Tina Comes. 2022. Machine learning for spatial analyses in urban areas: a scoping review. *Sustainable Cities and Society*, 85:104050.

YewKen Chia, Pengfei Hong, and Soujanya Poria. 2023. Flan-alpaca: Instruction tuning from humans and machines.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Eric M. Delmelle, Michael R. Desjardins, Paul Jung, Claudio Owusu, Yu Lan, Alexander Hohl, and Coline Dony. 2022. Uncertainty in geospatial health: challenges and opportunities ahead. *Annals of Epidemiology*, 65:15–30.

Vivek Doudagiri, Alakananda Vempala, and Eduardo Blanco. 2018. Annotating if the authors of a tweet are located at the locations they tweet about. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Xuke Hu, Zhiyong Zhou, Hao Li, Yingjie Hu, Fuqiang Gu, Jens Kersten, Hongchao Fan, and Friederike Klan. 2023. Location reference recognition from texts: A survey and comparison. *ACM Comput. Surv.* Just Accepted.

Diana Inkpen, Ji Liu, Atefeh Farzindar, Farzaneh Kazemi, and Diman Ghazi. 2017. Location detection and disambiguation from twitter messages. *Journal of Intelligent Information Systems*, 49(2):237–253.

Thiombiano Julie, Malo Sadouanouan, and Traore Yaya. 2023. A geolocation approach for tweets not explicitly georeferenced based on machine learning. In *Distributed Computing and Artificial Intelligence, 20th International Conference*, pages 223–231, Cham. Springer Nature Switzerland.

Rabindra Lamsal, Aaron Harwood, and Maria Rodriguez Read. 2022. Where did you tweet from? inferring the origin locations of tweets based on contextual information. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 3935–3944.

S. Teven Le et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, abs/2211.05100.

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models infer and disagree like humans? *ArXiv*, abs/2305.13788.

Chenliang Li and Aixin Sun. 2014. Fine-grained location extraction from tweets with temporal awareness. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, page 43–52, New York, NY, USA. Association for Computing Machinery.

Menglin Li and Kwan Hui Lim. 2022. Geotagging social media posts to landmarks using hierarchical bert (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12999–13000.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

S. Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Kateryna Lutsai and Christoph H. Lampert. 2023. Geolocation predicting of tweets using bert-based models. *ArXiv*, abs/2303.07865.

Chenyang Lyu, Jitao Xu, and Longyue Wang. 2023. New trends in machine translation using large language models: Case examples with chatgpt. *ArXiv*, abs/2305.01181.

Farshid Mirzaalian and Elizabeth Halpenny. 2019. Social media analytics in hospitality and tourism: A systematic literature review and future trends. *Journal of Hospitality and Tourism Technology*, 10(4):764–790.

Monika and Aruna Bhat. 2022. Automatic twitter crime prediction using hybrid wavelet convolutional neural network with world cup optimization. *International Journal of Pattern Recognition and Artificial Intelligence*, 36(05):2259005.

Aml Mostafa, Walaa Gad, Tamer Abdelkader, and Nagwa Badr. 2022. Pre-hlsa: Predicting home location for twitter users based on sentimental analysis. *Ain Shams Engineering Journal*, 13(1):101501.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020a. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020b. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

B. Rishi et al. 2021. On the opportunities and risks of foundation models. *ArXiv*.

Wei Shen, Yinan Liu, and Jianyong Wang. 2018. Predicting named entity location using twitter. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pages 161–172.

Rohit Shewale. 2024. Twitter statistics for 2024 — (facts after "x" rebranding).

Lihardo Faisal Simanjuntak, Rahmad Mahendra, and Evi Yulianti. 2022. We know you are living in bali: Location prediction of twitter users using bert language model. *Big Data and Cognitive Computing*, 6(3).

Yi Tay. 2023. A new open source flan 20b with ul2. `https://www.yitay.net/blog/flan-ul2-20b`. Accessed: 10/08/2023.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023a. UL2: unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won

Chung, Dara Bahri, Tal Schuster, Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023b. UL2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations*.

H. Touvron et al. 2023a. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023b. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Asahi Ushio, Francesco Barbieri, Vitor Sousa, Leonardo Neves, and Jose Camacho-Collados. 2022. Named entity recognition in Twitter: A dataset and analysis on short-term temporal shifts. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 309–319, Online only. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xinglei Wang, Meng Fang, Zichao Zeng, and Tao Cheng. 2023. Where would i go next? large language models as human mobility predictors. *ArXiv*, abs/2308.15197.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Fine-tuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhaomin Xiao and Eduardo Blanco. 2022. Are people located in the places they mention in their tweets? a multimodal approach. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2561–2571, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zhaomin Xiao, Yan Huang, and Eduardo Blanco. 2023. Context helps determine spatial knowledge from tweets. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 149–160, Nusa Dua, Bali. Association for Computational Linguistics.