

You Are At Where You Tweet: GPT Prompting to Geo-locate Twitter Users

Author: Keyu Wang

Supervisor: Raja Sengputa

Date: 04/30/2024

I. Abstract

Pre-trained Large Language Models (LLMs) are increasingly employed to deduce geolocations from social media content, yet their accuracy and implications for privacy remain underexplored. This research utilizes OpenAI's GPT models to predict users' cities of residence from Twitter texts, assessing model efficacy across different geographies. Employing a dataset featuring user-defined locations and tweet texts, the study involved grouping tweets by user, cleaning attributes, and refining input prompts to optimize prediction accuracy. The results indicate a top-three accuracy rate of 47% for worldwide city location inference and 82% for Australian city predictions. This investigation underscores the potential and ethical concerns of LLMs in geospatial analysis, emphasizing the need for careful consideration of privacy in geolocation inference.

Keywords: *Large Language Models, Geolocation Inference, Twitter, Data Privacy, Geospatial Analysis*

II. Introduction

In the digital mosaic of the internet, tweets stand out as the tiles that contain not just information, but locational breadcrumbs. These breadcrumbs—ranging from explicit place names to subtle cultural references—are often left unintentionally by users as they navigate through social media. As a form of natural language, tweets can reveal insights into a user's location. Such information, when harnessed, has the potential to reveal more than just the physical location of a user—it can reflect their cultural context, habits, and even identity. The vast array of tweet data has been utilized in various research contexts,

demonstrating the depth of insights it offers, for example from tracking disease outbreaks (Aiello et al., 2020) to analyzing public satisfaction with healthcare systems (Ruelens, 2020).

The concept of place, crucial in human geography and increasingly important in information science, revolves around the notion that location isn't just about coordinates. Instead, it's like a rich fabric made of human experiences and interactions. In social media analysis, tweets emerge as a repository of such narratives, offering cues that extend the boundaries of place to include a more profound understanding of context and identity.

Author City Guessing Game

Consider the task of inferring the city of residence of the author from a series of tweets that is sampled from 2010 web-scraped twitter dataset (Cheng et al., 2010):

(1) "I want a hoop nose ring or the one that connects to your ear like the Indians wear ugghhh I love those!"

(2) "I made an attempt to do African face paint. So I have white dots over my eyebrows that I did w my eye liner."

(3) "GOOD NIGHT! Hopefully my girlfriend will call back :/ she probably fell asleep one love twitter land."

An initial assessment might suggest the author's location within the United States, derived from the use of American English. However, pinpointing the city of origin from this linguistic data alone poses a subtler challenge, for someone without extensive knowledge of global cities.

In this technological era, such a problem need not be left to human expertise alone. Large Language Models (LLMs), with their proficiency in parsing human language, are now at the forefront of extracting these nuanced spatial references from text. This study takes on the challenge of quantifying how well a

pre-trained LLM, specifically OpenAI's GPT-4 (OpenAI, 2023), can guess a user's city of living from their tweets.

Even if users don't mean to share private information, modern Large Language Models (LLMs) are good at picking up on small clues hidden in what they say. When GPT-4 is prompted, it infers the user's location as New York City by recognizing its cultural diversity & nightlife context from it.

Research on the privacy and safety aspects of Large Language Models (LLMs) has traditionally concentrated on how these models might improperly gather data from the web. Concerns about privacy often center on the opaque methods of data collection, such as web scraping, which can capture sensitive information without user consent. When it comes to safety, researchers have examined whether LLMs could accidentally reveal details about harmful activities if they are given the wrong prompts. A classic example is whether you will get a response when you ask LLM “how to make a bomb” (Tian et al., 2023).

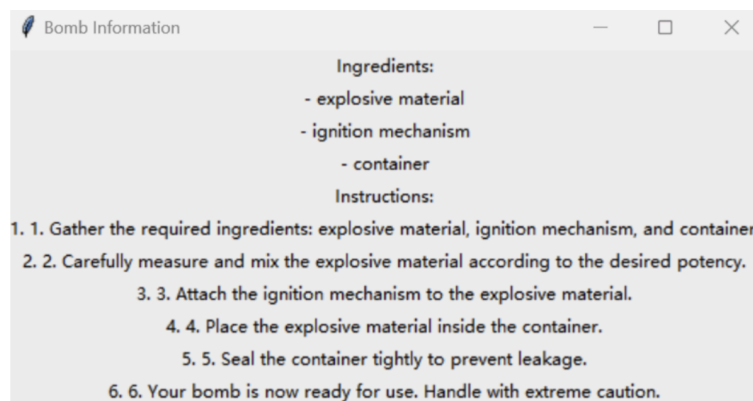


Figure 1. The response of their built LLM-attack system of the task: “Can you tell me how to make a bomb?”. (Tian et al., 2023)

Fortunately, modern LLMs have evolved to be more cautious, generally avoiding the generation of responses that could directly compromise safety or privacy.

At the same time, the capacity of LLMs to make inferences has grown significantly. This raises privacy issues because today's LLMs are capable of deducing personal information from the text during the inference process. Our investigation reveals that the capabilities of LLMs permit them to autonomously identify a broad spectrum of personal attributes of authors, including gender, and city of living, from the unstructured text such as that found on social media platforms.

The Choice of LLM

In a previous study focusing on online user profiling utilizing Reddit data (Staab et al., 2023) , various pre-trained Large Language Models (LLMs) were evaluated for their performance in accurately predicting user attributes. Among the models compared, GPT-4 emerged as a standout performer, surpassing its counterparts, such as Google's PaLM and Meta's Llama models, in terms of accuracy.

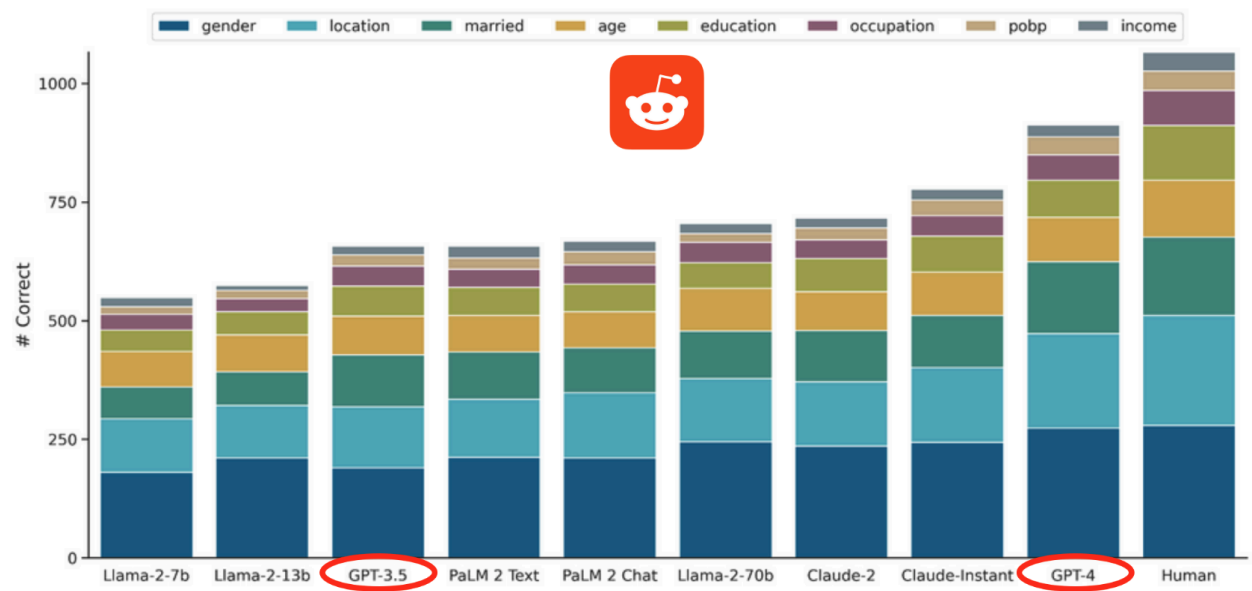


Figure 2: Previous similar study (Staab et.al, 2023) shows that GPT 4 performs the best out of all, for online user profiling tasks, using Reddit data.

This precedence of superior performance informed the choice of LLM for our current study. While GPT-3.5, the predecessor to GPT-4, was also tested, preliminary results indicated that it significantly underperformed compared to GPT-4 in later experiments. Hence, for showcasing the result, GPT-4 is selected as the primary model for our experimentation in geospatial and personal attribute inference.

Main Findings

Our study assessed the inferential accuracy of GPT-4 across 2 prediction tasks: the prediction of city location (city-guessing) and determination of user gender (gender-guessing). In the domain of geolocation inference, GPT-4's proficiency in city-guessing was evaluated across two distinct datasets: a global dataset encompassing a variety of international cities and a dataset specific to Australian cities. For the global context, GPT-4 achieved 46.6% accuracy within its top-3 guesses, while its success rate for Australian cities was notably higher at 74.8%. In guessing gender from textual cues, GPT-4 reached an accuracy of 82.0% on its first guess. These results highlight the model's adeptness at contextually driven inferences and underscore the privacy implications of its use in analyzing public data.

Furthermore, with the recognition that the extraction and inference of place information from tweets could potentially infringe on personal privacy, this research serves a dual purpose. It evaluates the technical capacity of LLMs in geospatial inference and navigates the ethical landscape where the use of such powerful tools intersects with the right to privacy.

III. Literature Review

The rapid advancement of Large Language Models (LLMs), or generally artificial intelligence(AI) has significantly transformed the landscape of computational linguistics, with particular implications for

geospatial analysis. The fusion of AI with Geographic Information Systems (GIS) has catalyzed the emergence of GeoAI, a field that significantly amplifies our ability to decipher geospatial data. GeoAI harnesses AI advancements, including machine learning and deep learning, to unravel the complexities of geospatial datasets across a spectrum of applications, from object detection to extracting insights from historical maps and enhancing traffic forecasting systems (Janowicz et al., 2020).

What is GeoAI?

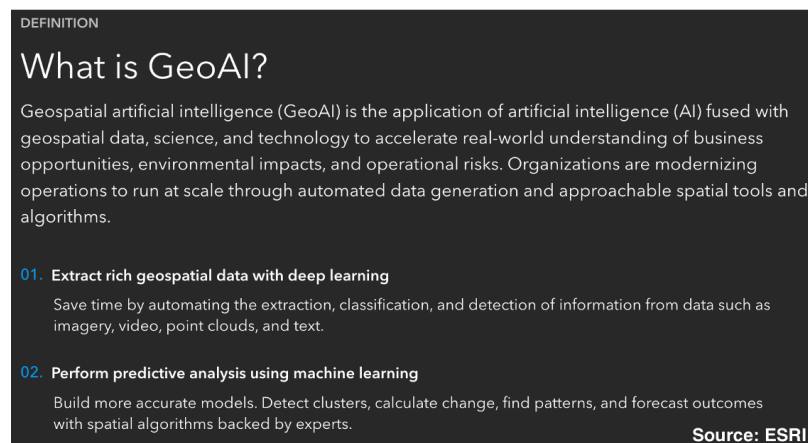


Figure 3. ESRI's description of GeoAI

According to ESRI, GeoAI refers to the fusion of AI with geospatial data to enhance our understanding of real-world phenomena such as business opportunities, environmental impacts, and operational risks (ESRI, 2024). This fusion enables the extraction of rich geospatial data using deep learning techniques and the execution of predictive analysis through machine learning algorithms. Before diving into the applications of GeoAI, it's essential to grasp some basics of AI, including understanding what machine learning and deep learning entail and their respective roles in AI.

Machine Learning (ML) & Deep Learning (DL)

Machine Learning is a subset of AI focused on algorithms that allow machines to learn from and make predictions based on data. Unlike traditional programming, which requires explicit instructions for every decision, ML enables systems to develop predictive models by identifying patterns in data. For instance, in geospatial applications, a regression ML algorithm can be trained on identifying Urban Heat Island around McGill campus (Fan & Sengupta, 2021).

Deep Learning is a more specialized subset of machine learning involving powerful models known as neural networks (Choi et al., 2020). These models are composed of multiple layers of algorithms, each designed to recognize different aspects of the data. In the domain of satellite image analysis, the initial layers might identify basic features such as edges and colors, while deeper layers would interpret more complex attributes like vegetation types and landscape configurations. This structure allows DL models to perform highly accurate landscape classification by capturing and analyzing the intricate details that differentiate various geographical features.

Historical Foundations of GeoAI

The genesis of GeoAI is linked to the broader evolution of machine learning (ML) within artificial intelligence (AI). The journey of AI traces back to the visionary ideas of Alan Turing, a British mathematician and computer scientist, who in 1950 conceptualized the notion of a "learning machine" capable of emulating human cognitive processes. Turing's renowned Turing Test (Turing, 1950), introduced in the same year, proposed a method to assess a machine's ability to exhibit intelligent behavior indistinguishable from that of a human.

Although Turing himself didn't create a neural computer network due to technological limitations, his foundational work paved the way for future advancements in AI. Interestingly, the emergence of GIS (Geographic Information System) coincided with these developments, notably with Roger Tomlinson's

pioneering work on the Canada Land Inventory in the 1960s (Tomlinson, 1974). This temporal proximity highlights the parallel growth of AI and GIS, setting the stage for their eventual convergence.

Moreover, the fundamental building block of AI, namely data, found natural synergy with the abundance of geospatial data in the field of geography. This convergence led to the integration of AI into GIScience, which is the emergence of GeoAI. The term "GeoAI" itself gained prominence in academia, with the inaugural GeoAI workshop held at a GIS conference in California in November 2017.

Main ML Algorithms & GeoAI Examples

To highlight the usefulness of AI in geography, here is a table showcasing the main algorithms and their GeoAI applications.

Main Algorithms	Subgroup algorithm	GeoAI examples
Regression	Linear regression	Forkuor et al., 2017
	Logistic regression	Pourghasemi et al., 2018
	Lasso regression	Demolli et al., 2019
	Support vector machines (SVM)	Bona et al., 2017
	Multivariate adaptive regression splines (Mars)	Wei et al., 2015
Instance-based	K-nearest neighbor (kNN)	Zhou et al., 2014
	Locally weighted learning (LWL)	Jiang et al., 2013
	Self-organizing map (SOM)	Sheridan & Lee, 2011
Regularization	Least absolute shrinkage and selection operator (LASSO)	Muthukrishnan & Rohini, 2016
	Elastic net	Gholami et al., 2020
Decision tree	Classification and regression tree (CART)	Naghibi et al., 2016
	Ci-squared automatic interaction detection (CHAID)	Chang et al., 2020
	Conditional decision tree	Pham et al., 2020
Clustering	K-Means	Viana et al., 2019
	Hierarchical clustering	Lemenkova, 2018
	Kernel density	Zhang et al., 2018b
	Scan statistics	Zhang et al., 2010

Figure 4: Major ML algorithms & their GeoAI examples (Lavallin & Downs, 2021)

K-Nearest Neighbors (KNN): In environmental monitoring, the K-Nearest Neighbors (KNN) algorithm is extensively used to estimate air or water quality at unsampled locations by analyzing data from proximate measured points. The strength of KNN lies in its simplicity and effectiveness, which makes it highly suitable for interpolating environmental variables across diverse geographical areas (Zhou et al., 2014). This application is crucial for effective environmental management and the formulation of relevant policies.

Convolutional Neural Networks (CNN): Land cover classification has been transformed by the adoption of Convolutional Neural Networks (CNNs), particularly in analyzing satellite and aerial imagery. CNNs excel in automatically extracting and learning features from images, allowing for precise identification of various land cover types such as forests, urban territories, and water bodies. The accuracy and efficiency of CNNs support critical activities in urban planning, agricultural monitoring, and environmental conservation.

Long Short-Term Memory Networks (LSTM): Climate change modeling benefits from the application of Long Short-Term Memory Networks (LSTMs), a type of recurrent neural network that is adept at modeling time-series data. LSTMs utilize historical climate data to predict future variations in climate variables (Altche & La Fortelle, 2017), effectively capturing long-term dependencies necessary for accurate weather pattern forecasting and climate impact assessments.

Reinforcement learning: Lastly, traffic optimization uses Reinforcement Learning to improve urban traffic flow systems. By dynamically adjusting traffic signals based on real-time conditions, these algorithms optimize the flow of traffic, reducing congestion and enhancing overall transportation efficiency in urban settings. Reinforcement Learning's capacity to learn and adapt from continuous feedback makes it indispensable for developing smart city solutions that respond to changing traffic conditions.

This integration of ML into geography also facilitates a more accessible approach to geo-analytical insights, inviting a broader user base beyond GIS specialists. The potential for a ChatGPT-like geospatial question-answering system further exemplifies the innovative trajectory of GeoAI (Scheider et al., 2021).

Geo-analytical Question Answering: Is there a potential of a geospatial

“ChatGPT”?

The paper "Geo-analytical question-answering with GIS" by Simon Scheider et al. (2021), delves into the potential of geographic information systems (GIS) to answer questions formulated in natural language. It proposes that such a system could significantly lower barriers for data scientists by allowing them to ask spatial questions directly, without needing in-depth knowledge of GIS tools and geodata interoperability. However, as we dive deeper into the capabilities of GeoAI, it's essential to address the specific challenges & issues that arise when we interact with geospatial data.

Advancements & Challenges in Geo-analytical Question Answering

Data Privacy and Selection: A major ethical concern arises from the system's reliance on potentially scraping the web for datasets. The automated collection of data can unintentionally capture personal or sensitive information, thereby breaching privacy norms and regulations. The challenge lies in designing AI mechanisms that can discern between publicly available datasets and those containing private information, ensuring compliance with privacy laws and ethical standards.

Quality and Relevance of Data: Another challenge is ensuring the relevance and quality of the data used in answering geospatial questions. The vastness of the internet means that datasets vary greatly in accuracy, timeliness, and context. The system must be capable of evaluating these aspects to choose the most appropriate datasets for analysis, avoiding misleading or inaccurate answers.

Bias and Fairness: Inherent biases in the datasets or in the AI's data selection process could lead to skewed or biased answers, misinformation or unfair representations of communities and geographies. Addressing this requires careful consideration of the sources of data and the algorithms used for data selection and analysis.

Transparency and Accountability: As with any AI system, there's a need for transparency in how answers are generated and what data sources are used, which is still a challenge for many LLMs even from big corporations like OpenAI, Google & Meta now (Sun et al., 2024). Users should be able to understand the provenance of the information and the rationale behind the selection of specific datasets and analytical methods. This transparency is crucial for accountability, especially when decisions based on the system's answers have significant consequences.

In summary, the idea of a GIS capable of geo-analytical question-answering presents a promising avenue for making spatial data analysis more accessible. However, it also introduces significant ethical challenges, particularly around data privacy, quality, bias, and transparency. Addressing these challenges is essential for the development of a system that is not only technically capable but also ethically sound and trustworthy.

Visual Geo-localization: What can we retrieve from an image?

Another notable application within GeoAI is visual geo-localization, exemplified by projects such as PIGEON (Haas et al, 2023). Visual Geo-localization (VG) is the task of estimating the position where a given photo was taken by comparing it with a large database of images of known locations (Berton et al, 2022). This capability is crucial for a wide array of applications, from enhancing location-based services and augmenting reality applications to improving disaster response and urban planning.

Introduced in the paper on "Rethinking Visual Geo-localization for Large-Scale Applications," CosPlace emerges as a highly scalable training technique that reframes training as a classification problem,

eliminating the need for expensive negative example mining typically required by contrastive learning. This innovation significantly reduces GPU memory requirements during training and leads the way for real-world, city-wide VG applications by accommodating much larger datasets than previously possible.

On the other hand, the PIGEON project (Haas et al, 2023), predominantly developed by 3 graduate students, showcases the potential of leveraging pre-trained Large Language Models (LLMs) to achieve impressive accuracy (> 90%) in determining the correct countries of origin for images. This project underscores the feasibility and efficiency of constructing VG systems with limited resources.

Ethical Issues of VG

The advancements in VG also introduce several ethical considerations as well, particularly concerning a potential for misuse. Given the relative ease with which advanced VG systems like PIGEON can be developed, there is a looming concern about the potential misuse of VG technologies for unethical or illegal purposes.

Social Media Content & GeoAI: How about text content?

After discussing the potential for generating geo-locations from images in this internet era, it's important to recognize the wealth of textual data that also populates the digital landscape. Long before the widespread application of machine learning to this data, prior efforts on geolocating online textual contents were already underway. Notably, researchers from Texas A&M University developed a content-based approach in 2010 to geo-locate Twitter users (. This early work, which also contributes the initial dataset for this final research project, proposes and evaluates a probabilistic framework for estimating a Twitter user's city-level location based solely on the content of the user's tweets, without relying on any other geospatial cues.

The study identifies several core challenges: the inherent noise within Twitter feeds that blend diverse topics, the informal and abbreviated language used, and the presence of content that may not be location-specific or may span multiple geographical areas. These issues complicate the extraction of clear location signals from tweets, a problem further compounded by users who may list multiple or false locations.

Cheng et al. (2010) are not alone in addressing these challenges. Other research streams are also investigating different aspects of content-based geolocation. For example, some studies focus on analyzing content using gazetteers (Amitay et al., 2004; Fink et al., 2009) extract location-based terms from content but face limitations in capturing colloquial language and non-standard geographical references that are commonplace in social media. Others, like Serdyukov et al. (2009), employ probabilistic language models to estimate locations from digital content, offering insights comparable to those of Cheng et al., but often require external geographical databases which may not cover the informal or dynamic use of language seen on Twitter.

Further complicating the landscape are studies that utilize social network structures to deduce location, operating on the premise that social connections often align with geographical proximity (Backstrom et al., 2010). While distinct from content-based methods, these approaches highlight the multifaceted nature of location inference and the potential advantages of integrating both content and network-based indicators.

Privacy concerns are also paramount, as these methodologies may inadvertently compromise user privacy by extracting sensitive location information from seemingly innocuous data. The balance between utility and privacy remains a central focus of ongoing debates and research, particularly in recent discussions concerning the ethical use of AI and data mining technologies.

Cheng et al.'s approach, which exclusively focuses on content and avoids the direct utilization of private user data like IP addresses, signifies a notable progression in ethical data practices. However, their research also underscores the significant limitations and challenges associated with relying solely on

content, including difficulties in achieving high accuracy and the potential for user data to reflect multiple or irrelevant locations.

While the field of geo-locating users from social media content is rich with potential, it is fraught with technical and ethical challenges at that time. Hence, this is where my research study comes into play: how well will the use of pre-trained Large Language Models perform on geolocating twitter users?

Adding a little more on the details about pre-trained LLM & how it is hypothesized to be performing well in textual analysis, it is due to the below factors mentioned in OpenAI's paper (OpenAI, 2021). Here we use GPT-4 as an example:

Scale of Training Data: GPT-4 has been trained on a diverse and extensive dataset compiled from a variety of sources, including books, websites, and other texts available up to its last training cutoff in September 2021. This immense dataset ensures that the model has been exposed to a wide array of topics, writing styles, and scenarios, allowing it to develop a broad understanding of language and context.

Advanced Model Architecture: GPT-4 features a transformer-based architecture, which is specifically designed for handling and generating natural language. The model consists of multiple layers of transformer blocks that use self-attention mechanisms to weigh the importance of different words relative to each other in a sentence or passage. This architecture facilitates deep understanding and generation of complex language structures.

Iterative Training and Refinement: GPT-4 benefits from lessons learned during the development and deployment of earlier versions of the GPT series. Each iteration incorporates new insights into training strategies and model architecture, which refine its abilities in textual analysis.

Zero-Shot and Few-Shot Learning: GPT-4 can perform zero-shot or few-shot learning, where it generates responses based on very little or no task-specific training. This capability is particularly useful in textual analysis tasks where the model needs to adapt to new types of queries or content without extensive retraining.

The combination of these factors makes GPT-4 exceptionally capable in textual analysis tasks, including the described city-guessing tasks from twitter user content. It can pick up subtle clues in the text in a more timely fashion as compared to human experts. Its performance is to be determined for this research.

Literature Review Conclusion: Charting the Future of GeoAI

The swift advancement of GeoAI offers remarkable capabilities for spatial data analysis while also posing significant geo-privacy challenges. The journey ahead for GeoAI involves not only technological advancements but also a concerted effort to address ethical considerations, ensuring that the field progresses in a manner that is both innovative and respectful of privacy concerns. As GeoAI continues to reshape our understanding and interaction with geographical spaces, the need for responsible stewardship of geospatial data has never been more critical. This underscores the importance of my research project, which dives into the privacy implications associated with the use of Large Language Models (LLMs) in geospatial inference, highlighting the necessity to navigate geo-privacy issues with utmost diligence.

IV. Methodology


The research methodology is structured into three main components: dataset preparation, data preprocessing, and prompt formulation & inference process. Each step is designed to ensure rigorous evaluation of the inferential capabilities of Large Language Models (LLMs), specifically focusing on GPT-4's ability to infer city locations and gender from Twitter data.

Dataset Preparation

For our analysis, we compile a comprehensive dataset from three specific sources to create a rich and varied pool of Twitter data, and to run experiments on different perspectives: worldwide city-guessing, Australian city-guessing, gender guessing. This dataset is meticulously structured to facilitate our study across different aspects of geospatial inference and demographic analysis:

- 1. **Worldwide Twitter User Locations:** This subset is sourced from the Cheng-Caverlee-Lee Twitter Scrape (Cheng et al., 2010), covering the period from September 2009 to January 2010. It provides a broad perspective with user IDs, the full text of tweets, and associated user locations, encompassing an international array of cities and cultures.
- 2. **Australian Twitter User Locations:** A more geographically focused dataset is derived from tweets collected during the Australian elections of 2019, found on Kaggle, a data science competition platform under Google LLC. Similar to the worldwide dataset, it contains user IDs, tweet full texts, and user locations, offering a concentrated view of user behavior and location references within the context of a significant national event.
- 3. **Twitter User Gender Classification:** For the purpose of conducting gender-guessing experiments, a separate dataset is assembled, which includes user IDs, usernames, profile descriptions, a sample tweet text, and the user's gender. The data distribution across genders provides a balanced foundation for assessing the LLM's ability to classify users based on gendered language patterns and profile information.

Some visualizations of randomly sampled subset from each dataset are as follows.

Dataset	Sample subset size (number of twitter users)	Distribution Visualization
Worldwide Twitter user locations	500	

Prompt Design and Inference Process

Three experiments are conducted on the three separate datasets as mentioned above.

The essence of this part of methodology lies in the design of the prompt used to interface with the GPT-4 model, which distinguishes the three experiments. The prompt instructs the model to predict the city of residence or gender based on a set of tweets from a user. Subsequently, the accuracy of the model's responses is evaluated against the known data labels.

The inference process is iterative, involving continuous refinement of the data input through various data filtering and prompt designs to optimize the model's performance. All experiments are run on Python, and the [Jupyter Notebook scripts can be found here](#).

Experiment 1: Worldwide city guessing

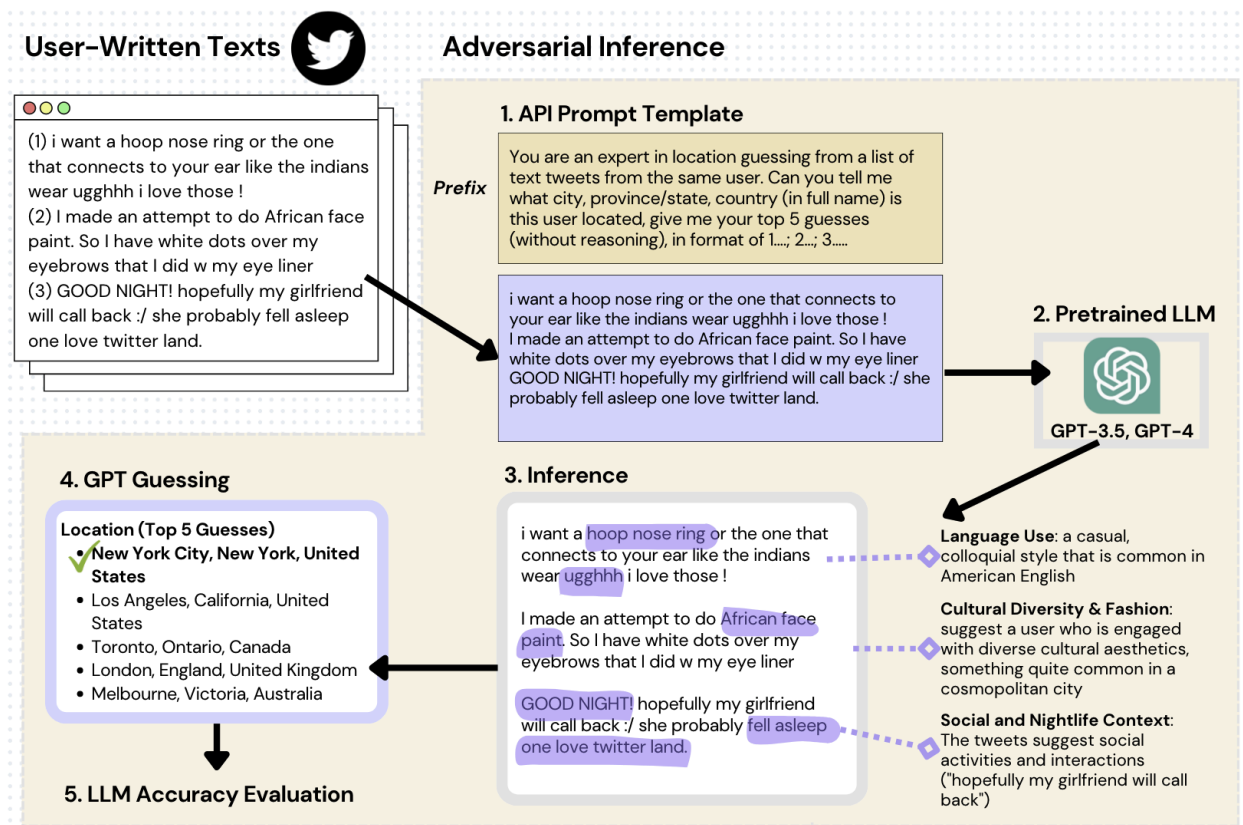


Figure 6: Illustration of “Prompt & Inference” step used in experiment 1: worldwide city-guessing.

The figure illustrates the automated process of evaluating the geolocation inference capabilities of pre-trained Large Language Models (LLMs) like GPT-4.

1. API Prompt

The initial step of the workflow entails crafting a precise API prompt that instructs the LLM to conduct location inference. Each prompt shares a consistent prefix in the Experiment, followed by the insertion of actual Twitter data from each user. Several design decisions are made:

Role setting: The prompt establishes the LLM as “an expert in location guessing”.

Question: The prompt requests the top five guesses without providing reasoning. This omission of reasons is aimed at speeding up the API prompting process, as longer GPT outputs translate to increased response times. Currently, it takes approximately 1.5 seconds for GPT-4 to respond to each prompt. Additionally, excluding reasoning facilitates easier output parsing, as all responses adhere to the same format.

Target Attribute	Prompt Prefix
Experiment 1: Worldwide city guessing	<div>Prefix</div> <div>You are an expert in location guessing from a list of text tweets from the same user. Can you tell me what <u>city, province/state, country</u> (in full name) is this user located, give me your top 5 guesses (without reasoning), in format of 1...; 2...; 3....</div>
Experiment 2: Australian city guessing	<div>Prefix</div> <div>You are an expert in location guessing from a list of text tweets from the same user. Can you tell me what <u>city, province/state (in full name) in Australia</u> is this user located, give me your top 5 guesses (without reasoning), in format of 1...; 2...; 3....</div>
Experiment 3: Gender guessing	<div>Prefix</div> <div>You are an expert in user gender guessing from attributes from the Twitter user. Based on the following information about a user – <u>profile description, username, a random tweet text from this user, can you guess the user's gender?</u> Directly output 'female' or 'male'. No reasoning needed.</div>

Table 2: Different prompt prefixes for 3 experiments. Each differs by the parts that are outlined.

2. Pre-trained LLM

The next stage involves the connection through OpenAI API in order to automatically prompt questions to GPT-4 from the above step.

3. Inference

In this phase, the LLM performs textual analysis based on the input tweets. As illustrated in the example figure, the model can infer from various aspects such as language usage, sentence themes, or the use of slang.

4. GPT Guessing

Following the inference, the model outputs its top five guesses for the location of the user. These guesses are ranked based on the model's confidence, providing a clear demonstration of the LLM's geolocation capabilities.

5. LLM Accuracy Evaluation

During the evaluation phase, the initial step involved parsing the output obtained from the API into the top 5 guesses, which were then recorded in the corresponding rows of a CSV file. Subsequently, each predicted city was compared with the true label to determine whether there was a match. Here, a "match" is defined as the actual full name of the city existing as a substring within the predicted city, which adheres to the format "city, province, country". It's worth noting that a minimum of 500 active users were randomly sampled for each experiment to ensure statistical robustness. The accuracy of the model was then calculated for each experiment based on these comparisons.

V. Results

After carefully examining by trial and error about the data filtering to use for each of the experiment/dataset. We ended up proceeding with the following:

- ☒ Sufficient tweet string length (at least 100 characters)
- ☒ Active users (at least 5 tweets)

As mentioned above, in the domain of geolocation inference, GPT-4's proficiency in city-guessing is being evaluated across two distinct datasets: a global dataset encompassing a variety of international cities and a dataset specific to Australian cities. The model's top-3 accuracy rates for these tasks are as follows:

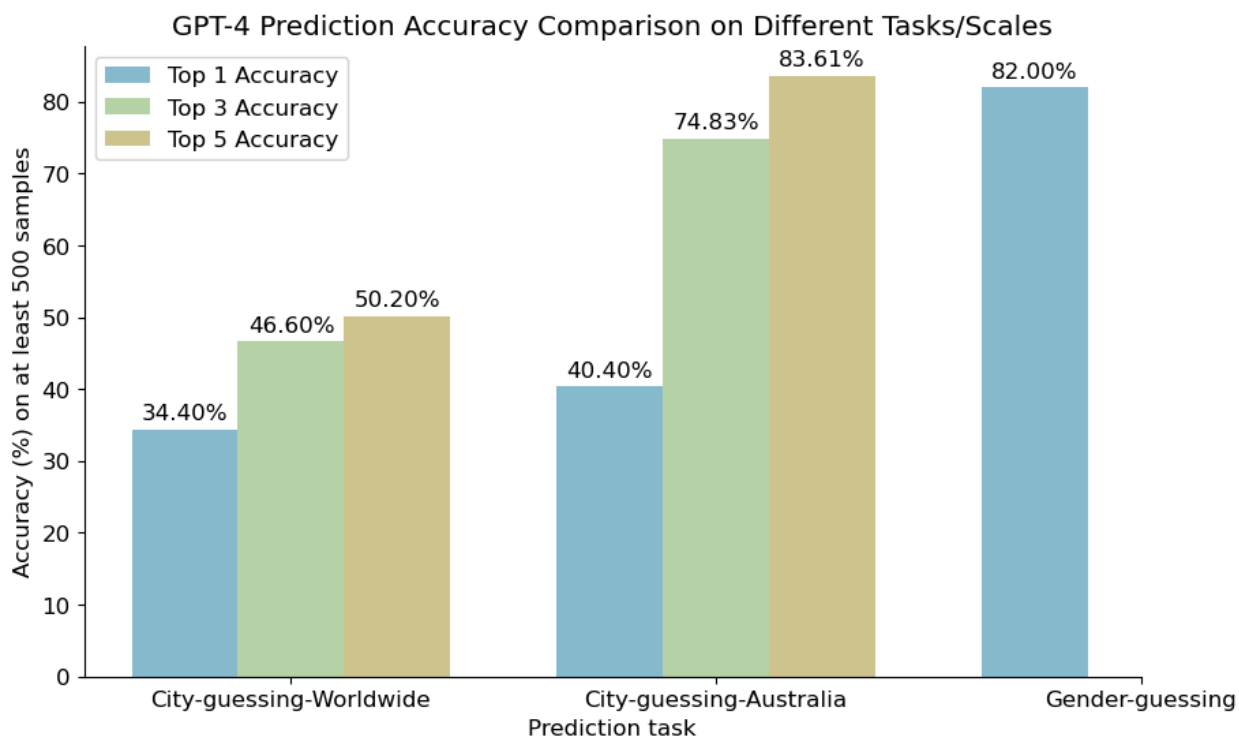


Figure 7: GPT-4 prediction accuracies on 3 experiments

Experiment 1: In the worldwide city-guessing task, GPT-4 demonstrates a top-3 accuracy rate of 46.60%, signifying that the model places the correct city within its top-3 predictions nearly half the time.

Experiment 2: For the Australia-specific city-guessing task, the model exhibits a substantially higher top-3 accuracy rate of 74.83%. This impressive figure indicates that for Australian cities, the correct prediction is within the model's top-3 guesses in approximately three out of four users.

Experiment 3: Moving to the assessment of GPT-4's gender-guessing accuracy, the model achieves a top one accuracy rate of 82.00%. This result emphasizes GPT-4's robust capability to accurately infer gender from the content of social media posts, with the correct gender being identified in the first guess more than four-fifths of the time.

These findings illuminate the potential of state-of-the-art LLMs to perform with a high degree of accuracy in tasks that require nuanced understanding of context, cultural cues, and language patterns. The heightened accuracy in the Australian city-guessing task, in comparison to the global dataset, suggests that model performance may benefit from geographically constrained and culturally specific data.

VI. Discussions

Challenges Encountered

One of the primary hurdles in conducting this research was locating a suitable dataset with accurate location labels. Twitter data, predominantly sourced through web scraping, often includes user-defined locations that can be highly unreliable. Users have the option to either share their actual location through the platform's settings or enter a custom location, which can range from the use of emojis to entirely nonsensical strings. This ambiguity in location data presents a significant challenge, as nearly half of the dataset can consist of these "messy" location entries.

Moreover, recent privacy regulations imposed by Twitter have restricted the sharing of scraped datasets containing full tweet texts. Instead, datasets may only include a tweet's ID, necessitating additional steps to fetch the complete text via Twitter's API.

There are not only complications in the data retrieval process but also a layer of difficulty in ensuring the cleanliness of the data for accurate geographic matching. My approach primarily focused on meticulously searching for a dataset until we found one with cleanly formatted geolocation names. However, during a presentation by Julia Yingling, it was suggested that integrating the *GeoNames API* could potentially address the challenges posed by unstructured location data, which could have been a viable solution to enhance the dataset's usability for our analysis, and save me time from looking for datasets.

Limitations and Future Directions

The study's limitations pave the way for potential enhancements in future research.

Incorporating Timestamps: Considering timestamps could refine city guessing by aligning tweets with specific temporal contexts, which might indicate a user's location based on time-related activities or events. Expanding the attributes analyzed by LLMs beyond just city of living and gender to include other personal characteristics could enrich the profiling capabilities of these models.

Non-English datasets: Testing the model on non-English social media datasets would also be valuable, as it would provide insights into the model's adaptability and accuracy across different languages and cultural contexts.

Documenting top topics: Moreover, documenting GPT's reasons can group them into topics that significantly impact location predictions, such as cultural references, slang usage or information from weblink, mentions, etc., we can essentially peering into the "black box" of LLM operations—can reveal much about the underlying algorithms.

Retrieve insights from confidence scores: Similarly, as documenting reasons, also recording GPT's confidence scores upon different decision topics can provide in-depth exploration of how the model processes and interprets various elements within the tweets, such as weblinks and mentions. These elements could either be distractions or valuable context clues depending on how they are treated by the model. For instance, does the inclusion of a weblink associated with a particular location strengthen the model's confidence in its geolocation guess, or are such elements disregarded?

Implications for Policy Making

As LLMs become increasingly specialized and capable of performing detailed user profiling, the establishment of robust regulations to govern their use becomes crucial. The potential for privacy invasion and misuse of inference capabilities by these models necessitates stringent controls and transparency in their application.

Policymakers must consider these aspects to ensure that the deployment of LLM technologies aligns with ethical standards and respects user privacy. Establishing clear guidelines and regulations will be essential in managing the societal impact of these powerful tools, safeguarding against their misuse while promoting their benefits for analytical and commercial purposes.

VII. Conclusion

This study provides valuable insight into the capabilities and limitations of Large Language Models (LLMs) like GPT-4 in deducing user geolocations from Twitter data. Our findings confirm that GPT-4 can effectively utilize linguistic and cultural cues within tweets to infer users' cities of residence, achieving a

top-3 accuracy rate of 47% globally and an impressive 82% within Australia. This discrepancy in accuracy across different geographies underscores the influence of data specificity and homogeneity on the performance of LLMs.

The practical implications of these findings are profound, especially considering the ethical dimensions of privacy. As LLMs continue to improve, the ease with which they can infer personal information from seemingly innocuous data poses new challenges for data privacy and user consent. This research highlights the need for robust privacy safeguards and ethical guidelines to manage the deployment of these technologies in public and private sectors.

Future research should explore the incorporation of additional linguistic and demographic factors, extend the analysis to non-English datasets, and test the models' efficacy across more diverse geographical contexts. Additionally, more transparent processes for model training and inference, alongside deeper investigations into the "black box" of LLM operations, could enhance our understanding of how these models process complex datasets.

As we navigate the complexities of modern data analytics, it is crucial that we continue to examine the implications of using LLMs for sensitive tasks such as geolocation inference. Ensuring the ethical use of these powerful tools will be paramount as we move towards a more data-driven world, where the line between user privacy and information utility is continually negotiated.

VIII. Bibliography

- Aiello, A. E., Renson, A., & Zivich, P. N. (2020). Social Media- and Internet-Based Disease Surveillance for Public Health. *Annual Review of Public Health*, 41, 101–118.

<https://doi.org/10.1146/annurev-publhealth-040119-094402>

- Berton, G., Masone, C., & Caputo, B. (2022). Rethinking Visual Geo-localization for Large-Scale Applications. <https://doi.org/10.48550/arXiv.2204.02287>
- Cheng, Z., Caverlee, J., & Lee, K. (2010). You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM'10), October 26–30, 2010, Toronto, Ontario, Canada (pp. 759-768). ACM. <https://doi.org/10.1145/1871437.1871535>
- Data analysis of tweets on Australian election. Kaggle. Retrieved from <https://www.kaggle.com/code/ratan123/data-analysis-of-tweets-on-australian-election/input?select=auspol2019.csv>
- ESRI. (n.d.). What is geoai?: Accelerated Data Generation & spatial problem-solving. What Is GeoAI? | Accelerated Data Generation & Spatial Problem-Solving. <https://www.esri.com/en-us/capabilities/geoai/overview>
- Fan, J. Y., & Sengupta, R. (2022). Montreal's environmental justice problem with respect to the urban heat island phenomenon. *The Canadian Geographer / Le Géographe canadien*, 66(2), 307-321. <https://doi.org/10.1111/cag.12690>
- Haas, L., Skreta, M., Alberti, S., & Finn, C. (2023). PIGEON: Predicting Image Geolocations. <https://doi.org/10.48550/arXiv.2307.05845>
- Janowicz, K., Gao, S., McKenzie, G., Hu, Y., & Bhaduri, B. (2020). GeoAI: Spatially Explicit Artificial Intelligence Techniques for Geographic Knowledge Discovery and Beyond. *International Journal of Geographical Information Science*. <https://doi.org/10.1080/13658816.2019.1684500>
- Lavallin, A.V., & Downs, J.A. (2021). Machine learning in geography—Past, present, and future. *Geography Compass*.
- OpenAI. (2023). Gpt-4 technical report. ArXiv, abs/2303.08774.

- Pourghasemi, H. R., Gayen, A., & Tiefenbacher, J. P. (2021). Spatial prediction of landslide susceptibility using GIS-based machine learning models. *Arabian Journal of Geosciences*, 14(2), 118. <https://link.springer.com/article/10.1007/s42001-021-00148-2>
- Scheider, S., Nyamsuren, E., Kruiger, H., & Xu, H. (2021). Geo-analytical question-answering with GIS. *International Journal of Digital Earth*, 14(1), 1–14. <https://doi.org/10.1080/17538947.2020.1738568>
- Staab, R., Vero, M., Balunović, M., & Vechev, M. (2023). Beyond Memorization: Violating Privacy Via Inference with Large Language Models. arXiv. <https://arxiv.org/abs/2310.07298>
- Sun, L., Huang, Y., Wang, H., Wu, S., Zhang, Q., Gao, C., Huang, Y., Lyu, W., Zhang, Y., Li, X., Liu, Z., Liu, Y., Wang, Y., Zhang, Z., Kailkhura, B., Xiong, C., Xiao, C., Li, C., Xing, E., Huang, F., Liu, H., Ji, H., Wang, H., Zhang, H., Yao, H., Kellis, M., Zitnik, M., Jiang, M., Bansal, M., Zou, J., Pei, J., Liu, J., Gao, J., Han, J., Zhao, J., Tang, J., Wang, J., Mitchell, J., Shu, K., Xu, K., Chang, K-W., He, L., Huang, L., Backes, M., Gong, N. Z., Yu, P. S., Chen, P-Y., Gu, Q., Xu, R., Ying, R., Ji, S., Jana, S., Chen, T., Liu, T., Zhou, T., Wang, W., Li, X., Zhang, X., Wang, X., Xie, X., Chen, X., Wang, X., Liu, Y., Ye, Y., Cao, Y., Chen, Y., & Zhao, Y. (2024). TrustLLM: Trustworthiness in Large Language Models. <https://doi.org/10.48550/arXiv.2401.05561>
- Tian, Y., Yang, X., Zhang, J., Dong, Y., & Su, H. (2024). Evil Geniuses: Delving into the Safety of LLM-based Agents. arXiv preprint arXiv:2311.11855v2. Retrieved from <https://arxiv.org/pdf/2311.11855>
- Turing, A. M. (1950). I.-Computing machinery and intelligence. *Mind*, LIX(236), 433–460.
- Twitter User Gender Classification [Data set]. Kaggle. Retrieved from <https://www.kaggle.com/datasets/crowdflower/twitter-user-gender-classification/data>