GeoLocator: a location-integrated large multimodal model for inferring geo-privacy

Yifan Yang
yyang295@usc.edu

Yixian Zhang
yaoqinse@gmail.com

Daoyang Li
daoyangl@usc.edu

Shuju Sun
shujusun@usc.edu

Junhong Duan
junhongd@usc.edu

Junzhou He
junzhouh@usc.edu

Qingyang Wu
qfrankwu@gmail.com

Hao Liu
hao.liu@helsinki.fi

December 14, 2023

Abstract

Geographic privacy or geo-privacy refers to the keeping private of one's geographic location, especially the restriction of geographical data maintained by personal electronic equipment. Geo-privacy is a crucial aspect of personal security, however often goes unnoticed in daily activities. With the surge in the use of Large Multimodal Models (LMM), such as GPT-4, for Open Source Intelligence (OSINT), the potential risks associated with geo-privacy breaches have intensified. This study develops a location-integrated GPT-4 based model named GeoLocator and designed four-dimensional experiments to demonstrate its capability in inferring and identifying the locational information of input imageries and/or social media contents. Our experiments reveal that GeoLocator generates specific geographic details with high accuracy and consequently embeds the risk of the model users exposing geospatial information to the public unintentionally, highlighting the thread of online data sharing, information gathering technologies and LLM on geo-privacy. We conclude with the broader implications of GeoLocator and our findings for individuals and the community at large, by emphasizing the urgency for enhanced awareness and protective measures against geo-privacy leakage in the era of advanced AI and widespread social media usage.

Keywords: geoprivacy, GPT-4, image comprehension, Large Multimodal Model (LMM), Open Source Intelligence (OSINT)

1 Introduction

In today's digital era, the silent leakage of personal information is a growing concern, with geographic privacy being a pivotal area of focus. Geographic privacy pertains to the protection and confidentiality of geographic information linked to individuals. It primarily encompasses safeguarding data that discloses an individual's geographic location, such as real-time whereabouts, historical movement patterns, or any location-specific information that can be traced back to them. The significance of geographic privacy is paramount. However, maintaining this privacy poses a significant challenge in the age of ubiquitous smartphones and social media platforms. While services like navigation, travel ticketing sites, and social media offer convenience, they simultaneously risk compromising our geographic privacy through potential surveillance, unauthorized data mining, and third-party misuse. This concern is further amplified by the existing legal framework's inability to keep pace with the rapidly evolving technologies that threaten geographic privacy.

We've discovered an easily overlooked way to give away our geographic privacy, and our everyday photos often contain a lot of geographic privacy information. For example, we post a photo on social media of ourselves clocking out of a particular ballpark. We can infer your geographic location from this photo. Specifically, which ballpark is in your photo and what is the specific address information for that ballpark. One photo is enough to give away your geographic privacy. With the rapid development of large multimodal models such as GPT-4, which is capable of extracting, interpreting, and inferring

geographic information from your published images, this geographic information is enough to expose your privacy.The ability that GPT-4 has to infer image photos poses a significant threat to geographic privacy. These models have the potential to reveal precise location details directly from geotagged images or indirectly through contextual analysis. The potential harms and implications are far-reaching and multifaceted, including identity theft, personal security breaches, and the risk of serious intrusions into an individual's private life.

Recognizing the potential threat posed by GPT-4 to geo-privacy, we developed GeoLocator, a tool integrating GPT-4 with geolocation function, and demonstrate its capability in inferring and identifying the locational information of input imageries and/or social media contents. In order to evaluate and compare the capability of regular search engines, GPT-4, and GeoLocator in perpetrating privacy attacks, we designed a series of experiments in four perspectives based on the input of various datasets, including Google Maps images, daytime/nighttime images, and social media posts. Our experiments reveal that GeoLocator generates specific geographic details with high accuracy and consequently embeds the risk of the model users exposing geospatial information to the public unintentionally, highlighting the thread of online data sharing, information gathering technologies and LLM on geo-privacy. We conclude with the broader implications of GeoLocator and our findings for individuals and the community at large, by emphasizing the urgency for enhanced awareness and protective measures against geo-privacy leakage in the era of advanced AI and widespread social media usage.

2 Related Work

We commence with providing a comprehensive overview of the capabilities of Large Multimodal Models (LMMs) in attacking geo-privacy, and the key milestones and innovative techniques that have shaped the evolution of LMMs.

LMMs introduction/definition The transformative emergence of the Transformer architecture [1] set a new precedent in the field, laying a robust foundation for contemporary large language models(LLMs). This breakthrough was followed by the development of pivotal models in Natural Language Processing (NLP), notably GPT [2] and Bidirectional Encoder Representations from Transformers [3]. More recently, with the development of computing power and advanced training techniques such as instruction tuning'[4, 5, 6] and reinforcement learning from human feedback(RLHF), large language models, such as ChatGPT[7] can achieve superior result in various downstream applications without the need for task-specific tuning. For example, LLMs excel in abstract summarization, producing meaningful overviews of text passages. This capability can be particularly beneficial in fields with vast amounts of text, like legal practice, academic research, and medicine, aiding in efficient navigation of dense information repositories[8, 9].Furthermore, LLMs has the ability to understand context and user intent has led to applications in customer service, personal assistance, and interactive educational tools[10, 11].

Concurrently, there's a notable trend towards integrating LLMs with vision-based models, heralding a new era of LMMs. This integration expands the range of tasks they can perform and aligning more closely with the multimodal nature of human cognition. LMMs differ from LLMs by processing and interpreting both textual and other types of data such as images. This advancement led to groundbreaking advancements in visual understanding and reasoning. For instance, the proprietary GPT-4 model [7], renowned for its illustrative abilities, and open-source models like Large Language and Vision Assistant [12], have demonstrated exceptional skill in blending textual and visual information. These models have shown proficiency in tasks ranging from generating website code from visual prompts [13] to recognizing complex details in image-rich contexts [14]. Their success illustrates not only the versatility of LMMs in handling multimodal data but also their potential in transforming tasks that require an intricate understanding of both visual and textual elements. We continue to explore existing work integrating large multimodal models with a variety of tasks and examine the application of artificial intelligence in geography.

LMMs applications LMMs have a very wide range of application capabilities. LMMs are the most cutting-edge technology that has been widely employed in diverse domains. In the medical field, Hou et al. found that the current multimodal models, GPT-4 and Bard could handle visual assinments. For example, GPT-4 successfully solved 96.7% of the visual problems, facing minimal difficulty with only one Parsons problem [10]. Yuan et al. found that LMMs can be applied to enhance various aspects of healthcare. Particularly, it highlights the crucial role of LMMs, investigating their ability

to process diverse data types like medical imaging and Electronic Health Records (EHRs) to augment diagnostic accuracy [8]. Fabian et al proposed a novel zero-shot species classification framework that leverages multimodal foundation models. This framework involves instruction tuning vision-language models to generate detailed visual descriptions of camera trap images, using terminology similar to that of experts [15]. Picard et al. evaluated GPT-4V, a vision language model, in engineering design tasks, demonstrating its capabilities and limitations. Their study provides foundational insights for the application of vision language models in engineering[16]. Warner et al. explored the shift in medical AI systems towards deep learning models, focusing on LMM's impact on medical image analysis and clinical decision support systems[17]. Oh et al. introduced a LMM for radiation therapy, integrating clinical text with images, demonstrating enhanced performance in breast cancer treatment, a first in such clinical text integration for oncology[18]. Microsoft delved into the capabilities of GPT-4Vision, highlighting its proficiency in video understanding, visual reasoning, and other areas. They underscored the substantial potential applications of this technology in various sectors, including industry, medical fields, auto-insurance, and image generation [19]. In summary, LMMs have showcased its strong and diverse application capabilities in solving visual problems in the aforementioned domains. Following these existing studies, we take LMMs as the baseline model to further develop our location-integrated model, GeoLocator.

Geography-related LLMs The use of LLMs in the spatial science domain has been relatively limited until quite recently. Earilier this year, Roberts et al. explored the geographical knowledge and reasoning skills of GPT-4 through a series of experiments, ranging from basic tasks like location estimation to complex applications like route planning and itinerary creation. Their study highlights GPT-4's capability in geospatial reasoning and its potential for diverse applications in geography-related fields[20]. Then Deng et al. developed K2, a specialized language model for geoscience, trained on a tailored corpus, showing enhanced performance in geoscience-specific tasks like question answering and knowledge reasoning, setting a new standard for domain-specific language models[21]. Li et al.introduced GeoLM, a language model integrating geospatial data with linguistic information, using geo-entity anchors and spatial coordinate embeddings for enhanced geo-entities understanding[22]. Hu et al.developed a method that combines geo-knowledge with GPT models for improved extraction of location descriptions from social media messages during disasters. This approach, using only 22 training examples, achieved over 40% improvement in accuracy compared to standard named entity recognition methods, significantly aiding in the rapid and efficient response to disaster scenarios[23]. Recently, Bhandari et al. assessed the geospatial knowledge and reasoning capabilities of LLMs, using experiments on geocoordinate prediction, geospatial preposition analysis, and multidimensional scaling, revealing their potential in geospatial reasoning tasks[24]. Extending from the existing research, we reformulated the regular LMMs to create our location-integrated model, GeoLocator, and tested out its capacity in inferring geospatial information and geo-privacy.

3 Procedure/method to develop the new tool – GeoLocator

Based on the GPT-4, we have developed a tool capable of inferring location information from images, which we have named GeoLocator(https://chat.openai.com/g/g-qxqvMb6YJ-geolocator). The GPT-4 is a large multimodal model (accepting image and text inputs and emitting text outputs) that, while not as capable as a human in many real-world scenarios, has demonstrated human-level performance on a variety of professional and academic benchmarks. GeoLocator is a customized version of the ChatGPT that we created.GeoLocator's strength is in using the powerful feature extraction and linguistic inference capabilities of large multimodal models to infer location information from images. At the same time, we developed GeoLocator with a large number of model commands built in to avoid transferring lengthy contexts each time.

Step 1: Basic ideas of GeoLocator design

GeoLocator is a kind of customized version of ChatGPT, and we do not need to code to implement GeoLocator's functionality. It is created by visiting https://chat.openai.com/gpts/editor. Creating one is as simple as starting a conversation, giving it instructions and additional knowledge, and then choosing what it can do, such as searching the web, making images, or analyzing data. Instructions are the key of GeoLocator. The process of creating instructions involves engineering skills, such as defining

the model's roles as being good at OSINT frameworks, criminology and geography. The purpose of the model is described in clear language and its task is to extract every detail from the photographs and come up with a sound analysis. In building the instructions for GeoLocator, we emphasized the need to focus on image details, EXIF data, traffic rules, human and physical geography, and unique regional clues. Step-by-step reasoning was used to improve accuracy.GeoLocator was built on GPT-4, which has the inherent advantage that GPT-4 can infer geographic information on its own, but the extent to which it can do so remains unknown. The GeoLocator we created, after enriching the directives, our GeoLocator has the potential to be much stronger in its ability to infer geographic information.

When built with instructions, our GeoLocator has a variety of features. It provides detailed place names, street names, and coordinates in the final location result. It performs an initial analysis using only street view images and refines its inferences using contextual information provided by the user. Users may provide misleading information that needs to be recognized and effectively differentiated. GeoLocator expresses itself visually and verbally, for example by recognizing and highlighting road signs and landmarks, and then performs a reasonably verbal analysis. It uses a code interpreter to pre-process images to deal with situations such as blurring or low resolution, including noise reduction, resolution enhancement, zooming or cropping. GeoLocator could utilize DALL-E image generation to obtain rich visual information, such as mapping spatial relationships (geographic manuscript style), topographic maps, or street maps. Finally, we merge the external search engine validation step into GeoLocator to confirm its conjectures. Using GeoLocator's conclusions, we call external APIs to draw maps that provide feedback on street location information.

4 Experimental design to test out GeoLoctor

We then designated a series of experiments to test out the capacity of GeoLocator in inferring location information and evaluating its modelling performance. We compared the results of geospatial / location information identified by three tools / platforms: Google search engine, GPT-4 and GeoLocator, based on images & texts. Such a comparison was implemented to different languages. We observed that GeoLocator has the strongest location inference ability across three tools mentioned above and can reach street level.

Given the size of our study, for all experiments, we judge the performance of the model solely based on whether it can correctly infer the location of the information given.

Step 2: Prepare input of data

In our study to evaluate the effectiveness of Google search engine, GPT-4, and GeoLocator across various image types, we gathered a diverse set of data sources. This included images from Google Maps, photographs taken by our research team, Google Images, and posts from social media.

Google Maps served as a resource for geographically diverse and detailed images, offering both reliability and recency. Our dataset (Figure 2) comprises 100 locations from Google Maps, including 50 iconic landmarks (e.g., the Status of Liberty in New York), 50 street views without obvious landmarks. Moreover, with the help of Google Maps, we could access different angel's street views. We selected 40 images of 20 different locations from two different angel. Additionally, authors captured 40 images of 20 specific location at separate times (day and night) to assess changes in environmental conditions. 10 images from Google Images were also chosen to assess the impact of language input on the results. 3 social media posts combined with text and images from the authors' personal accounts. The inclusion of social media images aimed to mimic real-life scenarios.

Data Source	Description	Number of images
Google Maps	Iconic landmarks	50 images
	Street view without obvious mark	50 images
	images of 20 locations from two different angel	20 sets (40 images)
Taken by research team	images of 20 locations at separate times (day and night)	20 sets (40 images)
Google Images	Images of 10 locations from China to assess the impact of language input	10 images
Posts from social media	Social media posts sent by research team member	3 posts

Figure 1: Data source and description

Step 3: Compare images' location inference ability among Google search engines, GPT-4, and GeoLoctor

In this part, we want to evaluate the location inference capabilities of research engines, GPT-4 and GeoLocator by sending photos without text prompts, thereby determining whether advanced artificial intelligence tools have better inference performance. This experiment highlights the potential of artificial intelligence tools like GPT-4 and GeoLocator in assisting with geographic location inference. In the experiment, we uploaded same images, range from iconic landmark to street view and day / night images of the same place, to Google search engine, GPT-4, and GeoLocator, and judge the performance of tools above based on their predictions' precision.

Step 4: Compare the location inference based on image and text instruction

Although GeoLocator already have high accuracy on location inference, in this experiment, we wonder whether it could perform better if we provide additional textual prompts or additional images. In this experiment, we evaluated whether GeoLocator could improve its performance in deducing the location of images by providing additional image perspectives or textual prompts. We selected 40 images taken from 20 different locations with different angel, and 10 images we used in step 3 with compared lower accuracy (accuracy of inference in Country level). We then compared GeoLocator's prediction results before and after applying these additional images or textual instructions.

step 5: Test out the impact of languages on inference results

With doing more experiments, we found that GeoLocator's performance varies with different input languages. Therefore, we designed this experiment to evaluate the impact of different languages on the model's inference results. During the experiment, we provided the model with images containing text prompts in different languages to GeoLocator and observed its predictions for the geographic location of the images. And we checked if the model's predictions vary with the change of input language.

step 6: Test out the GeoLoctor's performance on social media posts

In the final experiment, we focused on a more realistic scenario, social media posts. Social media posts usually contain more complex information such as emotions, what a used is doing, and some meaningless images and text for location inferences. We explored whether a finely tuned LMMs like GeoLocator could now inference the location of a place that social media posts show or describe.

5 Results

With the help of GeoLocator, we observed a high accuracy of the inference of location in most kinds of images mentioned in experiment. With the advancement of technology, especially the progress in image inference and geolocation techniques, it is becoming possible to infer where a photo was taken.

5.1 compare images' location inference ability among Google search engines, GPT-4, and GeoLoctor

This experiment highlights the potential of artificial intelligence tools like GPT-4 and GeoLocator. It is not only improving the speed and accuracy of inference but also expands the capabilities of users in conducting such tasks. We selected 10 representative images to show in Figure 4 in detail. To simplify our results, we grouped the results of our experiments into four geographical categories (Country, State, City/Town, Street). These were color-coded for inference accuracy, ranging from dark green for the most precise to light green for the least. The outcomes of our experiments are displayed in Figure 3. We consider inferences accurate to the street level as successful, and this criterion is used to evaluate the accuracy of three tools' inferences.

In Google search image, since the location inference of Google search engine is based on tags of images, it works well for deducing pictures of tourist attractions or iconic buildings with rich information, making it easy for search engines to infer these locations. But it is difficult to infer the location of street views because there is little information about various street views on the internet, making it hard to establish web links and thus unable to infer street views.

For street views image inference, GPT-4 can provide methods to predict the shooting location of images through its model used for image inference and analysis. It can analyze various factors such as landmarks, natural features, architectural styles, vegetation types, and weather conditions in the photos to complete the inference and prediction of image locations.

Outperforming GPT-4, GeoLocator through our set instructions, can infer specific location information of street view images and perform a more complete inference process that GPT-4, thereby achieving better performance. It can predict not only the country or city of the picture but even infer the name of the street or building where the image was taken without specific road signs or street names.

Image Type	Sample size	Google search engine	GPT-4	GeoLocator
Iconic landmark	50	88%	60%	94%
Street view	50	16%	18%	54%
Daytime image	20	25%	40%	70%
Nighttime image	20	10%	15%	35%

Figure 2: Results of inference accuracy

Images	Data Type	Google Search Engine	GPT-4	Geolocator	Distance* (miles)
	Street View	State	City/ Town	Street	*0.0034*
	Street View	City/ Town	Unknown**	City/ Town	*10*
	Street View	State	Unknown**	Country	*32.49*
	Street View	Country	Unknown**	Country	*126.42*
	Landmark	Street	Unknown**	Street	*0.0044*
	Landmark	City/ Town	Street	Street	*0.0019*
	Landmark	Unknown**	Country	City/ Town	*55.22*
	Landmark	Street	Street	Street	*0.0449*
	Street View	City/ Town	Unknown**	City/ Town	*2.62*
	Nighttime Image	Unknown**	Country	City/ Town	*3.28*

Figure 3: Results of compare images' location inference ability among Google search engines, GPT-4, and GeoLocator *It indicates the distance between the actual address from the image and the address inferred by the Geolocator. **It means that the Search Engine, GPT-4, or Geolocator is not able to find the address. *** Geographical Categories: Country, State, City/Town, and Street. A deeper background color indicates that it infers a more specific address in terms of the Geographical Category mentioned above.

5.2 compare the location inference based on image and text instruction

We evaluated whether the GeoLocator could improve its performance in inferring the location of street view images by providing additional image perspectives or textual prompt when the inference of location is not at street level.

To assess GeoLocator's proficiency in inferring locations from images captured at various angles of the same site, we conducted a series of ten experiments where each scenario could potentially deduce

a street-level location, using Taipei 101 as a representative example. It did not give us an expected answer at the first, and we gave the model another image taken from the same place but at a different angel. Then GeoLocator can successfully infer the place where the photo was taken. With additional information, GeoLocator could get more information to deduce the location further precisely in the images.

To evaluate the accuracy of location information identified by inputting additional text prompts, we test GeoLocator's performance. We chose to highlight the USC UPC case as an example, where each scenario could potentially deduce a street-level location as well. By uploading an image representing USC UPC, GeoLocator first observed the environment in the picture and infer it is in a temperate of Mediterranean climate area. The people in the picture were dressed lightly and casually, and the activities looked like they were happening on a campus. Based on these clues, GeoLocator analyzing that the picture was like shot at a university in the United States. For this unsatisfactory answer, we gave the model three prompts until it gave us a street level answer. In this situation with richer information, GeoLocator showed better performance that relying on image input alone. This multimodal approach enabled GeoLocator to infer and understand the content of images, including specific details like road signs and street names more effectively.

The advantage of this multimodal analysis is that it is not just a simple overpay of different modalities of information, but rather a deep understanding and analysis of the interrelationships between these diverse types of data, leading to more comprehensive and precise conclusions. In this situation with more information, GeoLocator performed better than when only inputting images, being able to infer more images with specific details like road signs and street names.

5.3 test out the impact of languages on inference results

In this experiment, we found different input will impact GeoLocator's reasoning procedure to affect the inference results.

During the experiment, we provided the GeoLocator the model with images containing text prompts in different languages and observed its reference for the geographical location of the images. The results showed that the model's reference indeed vary with the change of input language. For example, although the model conducted similar analyses when inputting different languages, their thought processes were different: when the input text was in English, the model prompted us with operations such as natural geographic, flora, architecture, and infrastructure; where as when the input text was in Chinese, the model prompted us with operations to refer the country, regional and zoning clues, focusing on very specific clues. Moreover, when asked in English, the model answered that the picture was from a park, which did not occur when asked in Chinese.

The phenomenon may be caused by multiple factors. Firstly, the bias in the language-related datasets the model may have been exposed to during training could lead the model to associate more with images related to the geographical location of users of a particular language when processing text in that language. Secondly, texts in different languages may imply specific cultural and geographic background information, which the model might use as clues to predict the image's location.

5.4 test out the GeoLoctor's performance on social media posts

The last experiment simulates the effect of GeoLocator on inferring the location of a picture taken in a more complex real-world scenario. Since social media posts always contain photos of the same location in different orientations, and some text describing the time or place, and there contain some useless information, this requires a higher level of ability for GeoLocator to synthesize a location. Additionally, we asked GeoLocator to generate a personal profile of the person who posts, based on both the image and text. This profile includes details such as the individual's location, age, and gender.

The experiment consists of 3 sub-experiments, comprising two tourist posts and one daily life post. The results show that GeoLocator can infer the location down to the city level and provide a detailed personal profile of the individuals posting. It can precisely pinpoint the location to the exact street or area for two out of the three posts. Additionally, the distances between GeoLocator's estimations and the actual locations in the images are all less than 100 miles, being 1.23 miles, 38.01 miles, and 0.27 mile, respectively.

6 Discussion

Large Multimodal Models (LMMs), exemplified by innovations like GPT-4, have ushered in a new era in the processing and interpretation of images, synergizing the analysis of textual and visual content. Despite their significant capabilities, the extent to which these models can extract detailed geographic information from street-view images remains an area of emerging research. In the evolving landscape of social networks, where individuals frequently upload personally relevant images, the advent of advanced LMMs like GPT-V has intensified concerns over privacy invasion, a long-standing issue now magnified by these technologies. This research delves into the application of LMMs within Open Source Intelligence (OSINT), highlighting the ease with which privacy can be compromised through the extraction of geographic information from images.

Employing the latest iteration of GPT-4 and its 'GPTs bot' feature, this study provides tailored instructions for the analysis of both textual and visual inputs, mirroring real-world scenarios prevalent in social media. The methodology encompasses a sequence of steps: initial photographic analysis, in-depth investigative research using search engines, geospatial mapping, address deduction, and linguistic adaptation, all aimed at refining communication effectiveness.

The experimental results demonstrate a high degree of accuracy in location recognition from a variety of images, including street views, buildings, and dimly lit scenes, facilitated by a customized ChatGPT tool. These findings underscore the capability of LMMs to interpret a substantial amount of geographical information, with predictions frequently aligning closely with actual locations, particularly in images with low brightness and minimal detail.

However, LMMs are not without limitations. The efficiency of geographic recognition by LMMs can fluctuate based on the language of input, suggesting a constraint in their multilingual capacities. Moreover, LMMs encounter challenges in recognizing less-documented locales, such as certain street views, due to limited online location data availability.

Future Actions

In the context of utilizing LMMs in future works, several key challenges and considerations emerge. Firstly, the complexity of instructions is a critical factor, as lengthy or convoluted instructions can impair GPTs' ability to process text, especially in the middle sections. This necessitates the development of concise, structured instructions to optimize model performance. Secondly, the reproducibility of experiments with LMMs demands methods to stabilize their outputs, possibly through standardized interaction protocols. Different activation modes are suggested to enhance efficiency, with complex analysis and graph generation tasks deferred to later dialogues, and a quick-start mode for immediate responses. The integration of specialized criminal investigation skills into LMMs presents a unique challenge, requiring the translation of empirical and traditionally oral skills into a format suitable for machine learning. Lastly, the integration of plugin systems into LMMs is not straightforward, indicating a gap in the models' ability to leverage external tools effectively. These points collectively highlight the complexities, potential, and necessary precautions in employing LMMs for OSINT.

Security vulnerabilities are the major risk we want to emphasize. LMMs can inadvertently reveal built information or private knowledge bases. The precision with which these models can analyze and interpret geospatial data opens avenues for misuse, including surveillance, stalking, or data theft. Ethically, the use of such technology should be governed by strict guidelines and legal frameworks to prevent misuse. In addition, LMMs can accurately pinpoint locations from images, including street views. This capability can lead to unintended consequences, such as revealing a person's whereabouts or personal habits, which raises serious concerns about the potential misuse of such information, especially in scenarios where user consent and awareness are lacking.

To balance between innovation and privacy, there's a critical need to balance these innovations with ethical considerations and privacy rights. Firstly, developing technologies that respect user privacy and incorporate ethical decision-making processes is crucial. This balance is not easy to achieve but is essential for responsible AI development. Secondly, there should be transparency in how these models are used and the kind of data they process. Users should be informed about the potential for their data to be analyzed in this way and given control over their participation. Additionally, there should be clear accountability mechanisms for the misuse of such technologies. Lastly, privacy-enhancing technologies, such as image encryption and anonymization techniques, should be prioritized. These technologies can help mitigate the risks associated with the detailed geographic information that LMMs can extract from images.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[2] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[4] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 27 730–27 744. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf

[5] Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Arunkumar, A. Ashok, A. S. Dhanasekaran, A. Naik, D. Stap, E. Pathak, G. Karamanolakis, H. G. Lai, I. Purohit, I. Mondal, J. Anderson, K. Kuznia, K. Doshi, M. Patel, K. K. Pal, M. Moradshahi, M. Parmar, M. Purohit, N. Varshney, P. R. Kaza, P. Verma, R. S. Puri, R. Karia, S. K. Sampat, S. Doshi, S. Mishra, S. Reddy, S. Patro, T. Dixit, X. Shen, C. Baral, Y. Choi, N. A. Smith, H. Hajishirzi, and D. Khashabi, "Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks," 2022.

[6] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi, "Self-instruct: Aligning language models with self-generated instructions," 2023.

[7] OpenAI, "Chatgpt," 2023. [Online]. Available: https://openai.com/chatgpt

[8] M. Yuan, P. Bao, J. Yuan, Y. Shen, Z. Chen, Y. Xie, J. Zhao, Y. Chen, L. Zhang, L. Shen, and B. Dong, "Large language models illuminate a progressive pathway to artificial healthcare assistant: A review," 2023.

[9] J. Holmes, S. Ye, Y. Li, S.-N. Wu, Z. Liu, Z. Wu, J. Hu, H. Zhao, X. Jiang, W. Liu, H. Wei, J. Zou, T. Liu, and Y. Shao, "Evaluating large language models in ophthalmology," 2023.

[10] I. Hou, O. Man, S. Mettille, S. Gutierrez, K. Angelikas, and S. MacNeil, "More robots are coming: Large multimodal models (chatgpt) can solve visually diverse images of parsons problems," 2023.

[11] "Driving a large language model revolution in customer service and support," https://www.databricks.com/blog/driving-large-language-model-revolution-customer-service-and-support, accessed: 2023-11-12.

[12] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023.

[13] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," 2023.

[14] Y. Zhang, R. Zhang, J. Gu, Y. Zhou, N. Lipka, D. Yang, and T. Sun, "Llavar: Enhanced visual instruction tuning for text-rich image understanding," 2023.

[15] Z. Fabian, Z. Miao, C. Li, Y. Zhang, Z. Liu, A. Hernández, A. Montes-Rojas, R. Escucha, L. Siabatto, A. Link, P. Arbeláez, R. Dodhia, and J. L. Ferres, "Multimodal foundation models for zero-shot animal species recognition in camera trap images," 2023.

[16] C. Picard, K. M. Edwards, A. C. Doris, B. Man, G. Giannone, M. F. Alam, and F. Ahmed, "From concept to manufacturing: Evaluating vision-language models for engineering design," 2023.

[17] E. Warner, J. Lee, W. Hsu, T. Syeda-Mahmood, C. Kahn, O. Gevaert, and A. Rao, "Multimodal machine learning in image-based and clinical biomedicine: Survey and prospects," 2023.

[18] Y. Oh, S. Park, H. K. Byun, J. S. Kim, and J. C. Ye, "Llm-driven multimodal target volume contouring in radiation oncology," 2023.

[19] Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, and L. Wang, "The dawn of lmms: Preliminary explorations with gpt-4v(ision)," 2023.

[20] J. Roberts, T. Lüddecke, S. Das, K. Han, and S. Albanie, "Gpt4geo: How a language model sees the world's geography," 2023.

[21] C. Deng, T. Zhang, Z. He, Y. Xu, Q. Chen, Y. Shi, L. Fu, W. Zhang, X. Wang, C. Zhou, Z. Lin, and J. He, "K2: A foundation language model for geoscience knowledge understanding and utilization," 2023.

[22] Z. Li, W. Zhou, Y.-Y. Chiang, and M. Chen, "Geolm: Empowering language models for geospatially grounded language understanding," 2023.

[23] Y. Hu, G. Mai, C. Cundy, K. Choi, N. Lao, W. Liu, G. Lakhanpal, R. Z. Zhou, and K. Joseph, "Geo-knowledge-guided gpt models improve the extraction of location descriptions from disaster-related social media messages," *International Journal of Geographical Information Science*, vol. 37, no. 11, p. 2289–2318, Oct. 2023. [Online]. Available: http://dx.doi.org/10.1080/13658816.2023.2266495

[24] P. Bhandari, A. Anastasopoulos, and D. Pfoser, "Are large language models geospatially knowledgeable?" 2023.