Geo-Foundation Models

Preprint · February 2024

1 author:

Gengchen Mai
University of Texas at Austin

INTERNATIONAL ENCYCLOPEDIA OF GEOGRAPHY: PEOPLE, THE EARTH, ENVIRONMENT, AND TECHNOLOGY

Geo-Foundation Models

Gengchen Mai[a,b]

[a] Spatially Explicit Artificial Intelligence Lab, Department of Geography, University of Georgia, Athens, Georgia, 30602, USA;
[b] School of Computing, University of Georgia, Athens, Georgia, 30602, USA

ARTICLE HISTORY
Compiled January 21, 2024

ABSTRACT
As a group of task-agnostic pre-trained large-scale neural network models that can be later adapted to numerous downstream tasks, foundation models have made a significant impact on academia, industry, and society. Meanwhile, several efforts have been made to develop foundation models for the geoscience domain. We call them Geo-Foundation Models (GeoFMs). In this paper, we address the "what," "why," and "how" of geo-foundation model development. The uniqueness of geographic data and tasks are highlighted and the necessary steps for GeoFM development are described in detail. This paper provides a general guideline for GeoFM research, and we advocate for a collaborative effort among academia, industry, and society to develop a reliable, sustainable, and ethically aware GeoFM framework.

KEYWORDS
Geospatial Artificial Intelligence; Large Language Models; Multimodal Foundation Models; Spatial Representation Learning

1. What are Foundation Models?

Foundation models (FMs) refer to a group of deep neural network models with a large number of learnable parameters that are pre-trained on **internet-scale datasets** in a **task-agnostic** manner and can be later **adapted to various downstream tasks** via model fine-tuning, few-shot learning, or even zero-shot learning (Bommasani *et al.* 2021, Mai *et al.* 2022a, 2023a). The concept of Foundation Models was proposed by a group of Computer Science faculty members at Stanford University to underscore these models' critically central yet incomplete character (Bommasani *et al.* 2021). The term "foundation" is so attractive but as Bommasani *et al.* (2021) said, "from a technical point of view, the idea of foundation models is not new." They are established based on techniques such as Transformer architecture (Vaswani *et al.* 2017), large-scale pre-training datasets such as CommonCrawl[1], unsupervised learning, self-supervised learning, and so on which have been established for years.

According to the data modalities these models can handle, we can roughly classify foundation models into three types: 1) language foundation models, or so-called large language models (LLMs), such as Bart (Lewis *et al.* 2019), PaLM (Wei *et al.* 2022),

[1] `https://commoncrawl.org/`

GPT-2 (Radford *et al.* 2019), GPT-3 (Brown *et al.* 2020), InstructGPT (Ouyang *et al.* 2022), ChatGPT (OpenAI 2022), OPT (Zhang *et al.* 2022), and LLaMA (Touvron *et al.* 2023); 2) vision foundation models such as DINO (Caron *et al.* 2021), SAM (Kirillov *et al.* 2023), and SegGPT (Wang *et al.* 2023b); and 3) multimodal foundation models such as CLIP (Radford *et al.* 2021), DALL-E 2 (Ramesh *et al.* 2022), Stable Diffusion (Rombach *et al.* 2022), KOSMOS-1 (Huang *et al.* 2023b), KOSMOS-2 (Peng *et al.* 2023), LLaVA (Liu *et al.* 2023a), GPT-4 (OpenAI 2023), and Gemini (Team *et al.* 2023). Different foundation models are capable of performing different sets of downstream tasks. For instance, most LLMs can perform various natural language processing tasks such as reading comprehension, question answering, sentence classification, machine translation, text summarization, and so on. Vision FMs such as SAM and DINO were developed to handle multiple vision tasks such as different segmentation tasks. However, both LLMs and Vision FMs can only handle one data modality which significantly limits the models' generalization ability. Nowadays, multimodal foundation model development is raised as the new frontier of foundation model research (Mai *et al.* 2023a, OpenAI 2023).

After OpenAI introduced ChatGPT on Nov. 30th, 2022, we have witnessed a huge impact of such technique not only on the traditional computer science domains such as natural language processing (NLP), computer vision (CV), and artificial intelligence (AI), but also on numerous other domains such as medicine (Moor *et al.* 2023, Li *et al.* 2023a), biology (Ma and Wang 2023), chemistry (Horawalavithana *et al.* 2022), agriculture (Lu *et al.* 2023), education (Latif *et al.* 2023, Lee *et al.* 2023), art and humanities (Liu *et al.* 2023b), and geography (Mai *et al.* 2023a, Nguyen *et al.* 2023, Roberts *et al.* 2023, Zhang *et al.* 2023b, Hu *et al.* 2023b, Jakubik *et al.* 2023, Lacoste *et al.* 2023, Manvi *et al.* 2024, Balsebre *et al.* 2023, Xie *et al.* 2023, Rao *et al.* 2023a, Tan *et al.* 2023). In the geography and the general geoscience domain, with the popularity of FMs, Geo-Foundation Models have been quickly raised as a new research direction.

2. What are Geo-Foundation Models?

Geo-Foundation Models (GeoFMs) (Mai *et al.* 2023c, Xie *et al.* 2023, Rao *et al.* 2023a) are a subset of foundation models that focus on explicit representations of spatial primitives such as spatial interaction, spatial stationarity, spatial heterogeneity, and so forth, and encode rich information about places and regions (Janowicz *et al.* 2020, Li *et al.* 2021, Mai *et al.* 2022b). After large-scale model pre-training, these Geo-Foundation Models are expected to match or even outperform the state-of-the-art **task-specific** GeoAI models on the corresponding downstream geospatial task by rather simple model adaption via zero-shot, few-shot learning, or model fine-tuning.

Although we are still at the early stage of GeoFM research and most existing research is mainly focusing on investigating the effectiveness of existing foundation models on various geospatial tasks (Mai *et al.* 2023a, Lu *et al.* 2023, Roberts *et al.* 2023, Zhang *et al.* 2023b,a, Hu *et al.* 2023b, Manvi *et al.* 2024, Tan *et al.* 2023) via zero-shot and few-shot in-context learning by using prompt engineering, we start to see some major efforts to develop specialized foundation models for different geospatial tasks. Most of these efforts follow the practice of the general-purpose foundation models but incorporate spatial thinking into the model architecture design and pre-training dataset construction.

One example is ClimaX (Nguyen *et al.* 2023), a weather and climate foundation model, which can be adapted to multiple weather and climate tasks such as short-term, middle-term, and long-term weather forecasting, climate projection, and climate downscaling. Nguyen *et al.* (2023) showed that ClimaX can outperform the state-of-the-art task-specific models on respective tasks such as the operational Integrated Forecasting System (IFS) (Wedi *et al.* 2015) for weather forecasting task. Another example is Prithvi (Jakubik *et al.* 2023), a transformer-based GeoFM. After pre-training Prithvi on 1TB of multispectral satellite imagery from the Harmonized Landsat-Sentinel 2 (HLS) dataset, Jakubik *et al.* (2023) demonstrated the effectiveness and generalization ability of Prithvi on multiple satellite-based geospatial tasks including multi-temporal cloud gap imputation, flood mapping, wildfire scar segmentation, and multi-temporal crop segmentation via task-specific model fine-tuning. Although ClimaX and Prithvi focus on completely different sets of geospatial tasks, both of them are based on vision foundation models. More specifically, both model architectures are modified based on Vision Transformer (ViT) (Dosovitskiy *et al.* 2020) by incorporating some spatial thinking into model design (Janowicz *et al.* 2020). ClimaX modifies ViT to accommodate spatially incomplete climate data and uses a cross-attention mechanism to aggregate different climate variable data at the same spatial location. Prithvi modifies the patch position encoding component of ViT into a 3D position encoding module in order to jointly consider the spatial and temporal dimensions of satellite data.

Other than GeoFMs derived from vision FMs, we also see major efforts in developing geoscience-specific large language models such as K2 (Deng *et al.* 2023). K2 geoscience large language model (Deng *et al.* 2023) was established by adapting a pre-trained general-domain LLM, LLaMA-7B model (Touvron *et al.* 2023), on geoscience text corpus. Similar to the training recipe of a general-domain LLM, Deng *et al.* (2023) utilized a two-step strategy to obtain K2. The first step is to do further unsupervised pre-training on a geoscience text corpus which consists of 1 million pieces of geoscience literature including geoscience-related Wikipedia pages, geoscience papers' abstracts, and open-access geosciences papers published in selected high-quality geoscience journals. The resulting LLM, named GeoLLaMA-7B, underwent additional supervised finetuning (the second step) on a geoscience instruction tuning dataset called GeoSignal by using a lightweight training technique called LoRA (Hu *et al.* 2021) which can significantly reduce the number of trainable parameters and speed up LLM training process. The effectiveness of K2 LLM was demonstrated on eight NLP tasks that require geoscience knowledge such as explanation, named entity recognition (i.e., place name recognition (Wang *et al.* 2020, Hu *et al.* 2023a) in the geoscience context), reasoning, fact verification, summarization, text classification, word semantics, and question answering. Although the results are promising, it's notable that K2's downstream tasks are classic NLP tasks requiring geoscience knowledge, rather than the geospatial semantic tasks commonly tackled by GIScientists. For example, the reasoning task considered by K2 mainly focuses on co-occurrence patterns of geo-entities in text paragraphs but not spatial reasoning (Randell *et al.* 1992, Regalia *et al.* 2019, Zhu *et al.* 2022, Mai *et al.* 2023b) and temporal reasoning (Renz and Nebel 2007, Cai *et al.* 2023) task which require explicit or implicit spatial or temporal computations. However, we still consider K2 as a major advancement of Geo-Foundation Model research.

Other than developing GeoFMs that can only handle single data modality as we discussed above, there are efforts aiming at developing multimodal GeoFMs. For example, inspired by the CLIP image-text pre-training framework (Radford *et al.* 2021), the contrastive spatial pre-training (CSP) framework proposed by Mai *et al.* (2023c) demonstrates one possible way to develop GeoFMs that can seamlessly handle image data, location data (i.e., geographic coordinates), or possible text data in a unified framework. By using geo-tagged images such as satellite images, or ground-level images with location metadata, CSP utilizes a dual-encoder framework to encode an image and its associated geographic coordinates with an image encoder (He *et al.* 2016, Dosovitskiy *et al.* 2020) and location encoder (Mac Aodha *et al.* 2019, Mai *et al.* 2020b, 2022c, 2023f, Cole *et al.* 2023) respectively. The resulting image embedding and location embedding are pre-trained in a contrastive learning objective. Mai *et al.* (2023c) demonstrated that this CSP pre-training is effective for downstream applications such as satellite image classification and species fine-grained recognition both in a few-shot learning and fully supervised learning setting. Although CSP itself is not a multimodal GeoFM, it lays a solid foundation for multimodal GeoFM pre-training (Mai *et al.* 2023a). Similar geo-aware self-supervised pre-training strategies can also be seen in GeoCLIP (Cepeda *et al.* 2023) and SatCLIP (Klemmer *et al.* 2023).

Another example of multimodal GeoFMs is CityFM (Balsebre *et al.* 2023), a self-supervised framework to train a GeoFM based on open available geographic data within a given geographical area of interest. More specifically, Balsebre *et al.* (2023) utilized nodes (e.g., points of interest), ways (e.g., roads, bridges, rivers, and building polygons), relations (e.g., a bus loop consisting of a set of polylines and points representing its paths and bus stops) as well as textual tags associated with these geographic entities from OpenStreetMap (OSM) to train a multimodal GeoFM for Singapore. In order to jointly consider different data modalities from OSM, they used three contrastive objectives: 1) Text-based contrastive objective: a BERT-based text encoder (Kenton and Toutanova 2019) is used to encode textual tags of each geographic entity into a text embedding which is contrastively learned against the spatially nearby entities' text embeddings. 2) Vision-language contrastive objective: a shape embedding and size embedding of a building polygon are computed based on a ResNet18-based (He *et al.* 2016) image encoder and a multilayer perceptron (MLP) respectively. The visual embedding of this building is then computed as the arithmetic mean of its shape and size embeddings which is later contrastively learned against this building's text embedding from the first objective. 3) Road-based contrastive objective: a road segment embedding is contrastively learned against other road segment embeddings such that two road segments that are traversed by similar numbers of bus loops will have similar embeddings. The effectiveness of CityFM was demonstrated on three distinct urban data science tasks: traffic speed inference, building functionality classification, and urban region population estimation.

All these GeoFMs we discussed above highlight the uniqueness and difficulties of geospatial problems. But why do we need GeoFMs in the first place? In the following, we will discuss the necessity of GeoFMs.

3. Why do we need Geo-Foundation Models?

Two arguments can challenge the necessity of GeoFM development: **Q1** – If we have a general-domain FM that is supposed to cover the geoscience domain, why do we need to develop a domain-specific FM such as GeoFM? **Q2** – We have many widely used tasks-specific GeoAI models for different geospatial tasks, why do we need a foundation model for all these tasks? The answers to both questions are rooted in the uniqueness of geographic data and tasks.

3.1. *Unique data modalities*

One strong argument to contradict **Q1** is that *geospatial tasks will usually require unique data modalities that are rarely considered in current general-domain FM development* such as geospatial vector data, remote sensing images, geographic knowledge graphs, etc. Although some general-domain FMs such as CLIP (Radford *et al.* 2021) also consider satellite images, as far as we know, none of the existing general-domain FMs can handle geospatial vector data such as points, polylines, and polygons which are core data types used in almost all geospatial tasks.

Fortunately, we started to see some efforts which conducted self-supervised learning on geospatial vector data along with other data modalities such as CityFM (Balsebre *et al.* 2023), CSP (Mai *et al.* 2023c), GeoCLIP (Cepeda *et al.* 2023), SatCLIP (Klemmer *et al.* 2023), and GeoLM (Li *et al.* 2023b). The key to their success is so-called *spatial representation learning* (Mai *et al.* 2023e) which aims at developing representation learning models that can encode geospatial vector data into neural network embedding space. According to the geospatial vector data types, such representation learning models can be classified into location encoder (Mac Aodha *et al.* 2019, Mai *et al.* 2020b, 2022c, 2023f,c, Cole *et al.* 2023), polyline encoder (Rao *et al.* 2020, Soni and Boddhu 2022, Ha and Eck 2018, Yu and Chen 2022, Rao *et al.* 2023b), and polygon encoder (Veer *et al.* 2018, Jiang *et al.* 2019, Yan *et al.* 2021, Mai *et al.* 2023b, Yue *et al.* 2023). After representing geospatial vector data in the neural network embedding space, it is possible to integrate this new modality into the current FM framework by utilizing many popular multimodal FM pre-training methods such as InfoNCE-based contrastive learning (Oord *et al.* 2018). We believe this is one of the major future research directions for GeoFM research.

3.2. *Generalize geographic knowledge to different tasks, geographic regions, and temporal scopes*

In terms of **Q2**, a counterargument against it is that the generalization ability baked in the nature of all foundation models is particularly crucial for geography research, especially when considering the perspective of nomothetic geography (Schaefer 1953), or so-called scientific or theoretical geography, which aims at searching for general laws and principles that apply and are replicable everywhere and presumably at all times (Goodchild and Li 2021).

In fact, the necessity for model generalization across various geospatial tasks is underscored by the sheer volume of tasks within the geography domain. As a domain where foundation models have experienced substantial advancement, natural language processing has a well-defined set of tasks such as text classification, named entity recognition, reading comprehension, sentiment analysis, information retrieval, machine translation, question answering, and so on (Kenton and Toutanova 2019, Brown *et al.* 2020). This lays a solid foundation for FM development. In contrast, the field of geography, despite its plethora of tasks, lacks a universally accepted set of tasks. Example tasks that have been widely studied in GeoAI literature are street view image recognition and segmentation (Zhang *et al.* 2018, 2019, Kang *et al.* 2021, Lee *et al.* 2021), remote sensing image classification and segmentation (Jean *et al.* 2019, Ayush *et al.* 2021, Manas *et al.* 2021, Cong *et al.* 2022, Fuller *et al.* 2023, Mai *et al.* 2023c), satellite image super-resolution (Müller *et al.* 2020, Mei *et al.* 2020, He *et al.* 2021, Mai *et al.* 2023d), population density estimation (Manvi *et al.* 2024, Balsebre *et al.* 2023), socioeconomic index prediction (Yeh *et al.* 2020, Elmustafa *et al.* 2022, Manvi *et al.*

2024), trajectory synthesis (Rao *et al.* 2023b, Rempe *et al.* 2023), building pattern recognition (Yan *et al.* 2019), place name recognition and disambiguation (Mai *et al.* 2023a, Hu *et al.* 2023b), geographic question answering (Mishra *et al.* 2010, Chen *et al.* 2013, Mai *et al.* 2018, 2020c, 2021, 2020a, Lobry *et al.* 2020, Huang *et al.* 2019, Lobry *et al.* 2021, Scheider *et al.* 2021) to name a few.

Traditionally, each geospatial task has been individually studied and tackled by different task-specific GeoAI models. However, there exists a common problem in many geospatial tasks – *labeled geospatial datasets have very limited size and imbalanced geographic and temporal distributions*. For example, compared with many large-scale natural image classification (e.g., ImageNet (Deng *et al.* 2009)) and object detection (e.g., Microsoft COCO (Lin *et al.* 2014)) datasets, satellite image classification (e.g., BigEarthNet (Sumbul *et al.* 2019), UC Merced Land Use (Yang and Newsam 2010), and EuroSAT (Helber *et al.* 2019)) and object detection datasets (e.g., xView (Lam *et al.* 2018), NEON Tree Crowns Dataset (Weinstein *et al.* 2021)) usually have limited size and are restricted in certain geographic regions and temporal periods. This significantly limits these models' generalization ability to other geographic regions, temporal scopes, or slightly different task setups.

Geo-foundation models are one great way to overcome these challenges. While labeled geographic data are limited in size and spatiotemporal coverage, we can first pre-train a GeoFM on the massive unlabeled geospatial dataset which has a much larger size, better spatiotemporal coverage, and a low cost to collect. This pre-trained GeoFM is more robust to the geographic and temporal bias (Henry Wai-Chung 2001, Liu *et al.* 2022, Faisal and Anastasopoulos 2022) that exists in the labeled datasets and can be adapted to multiple geospatial downstream tasks (Mai *et al.* 2023c). This can significantly lower the effort of GeoAI model development and dataset construction, especially in some tasks where data are scarce or expensive to collect.

4. How to develop Geo-Foundation Models?

After highlighting the significance of GeoFM, we will now delve into the necessary steps required for the development of GeoFM.

4.1. *Define a comprehensive set of core GeoAI tasks*

As we said in Section 3.2, geography lacks a universally accepted set of tasks. So the first step towards GeoFM development is to define a comprehensive set of core GeoAI tasks as well as a task classification schema based on their underlying data modalities, output data types, spatial scale, spatial coverage, temporal scale, temporal coverage, etc. Since all geospatial tasks can be translated into geographic questions. So the geographic question classification schema proposed by Mai *et al.* (2021) can provide a general guide for this GeoAI task classification schema. Moreover, the spatial core concepts (Kuhn 2012) and the geo-analytical question answering framework (Scheider *et al.* 2021) can serve as the theoretical foundation for this classification. The resulting GeoAI task set and classification schema can serve as the foundation for any GeoFM development:

(1) **Define the task scope of a GeoFM**: currently, no existing foundation model is capable of adapting to every conceivable task. All FMs are targeted at a finite set of tasks. For example, LLMs such as LLaMA 2 and GPT-3 focus on all

kinds of natural language understanding tasks. SAM (Kirillov *et al.* 2023) can only handle different image segmentation tasks. Stable Diffusion and DALL-E target text prompt-based image generation tasks. Similarly, if we have a set of core GeoAI tasks, GeoFM developers can decide which task subset the resulting GeoFM can be adapted to.

(2) **Decide the necessary data modalities the GeoFM can handle**: after deciding the task scope, based on the setups of these selected GeoAI tasks, we can determine the necessary data modalities the resulting GeoFM should handle. The data modalities commonly used by different GeoAI tasks include satellite images, radar point clouds, geospatial vector data, street view images, geo-text data, geographic knowledge graphs, and so on. For example, if we hope the resulting GeoFM can tackle the geocoding task, the GeoFM should be able to handle text data, place hierarchy information, and geographic coordinates.

(3) **Determine the spatiotemporal scales and coverage that the GeoFM is capable of handling**: The selected GeoAI tasks can also help determine the spatiotemporal scale and coverage a GeoFM should be able to handle. For example, as a city-scale foundation model, CityFM (Balsebre *et al.* 2023) certainly cannot perform the image/text geolocalization task which requires geographic knowledge all over the world. ClimaX (Nguyen *et al.* 2023) will struggle to recover the climate scenario in the 10th century since it was pre-trained only on global projections of climate scenarios from 1850 to 2015.

4.2. *Standing on the shoulders of giants*

One thing we need to keep in mind is that we should avoid "reinventing the wheels" and try to "stand on the shoulders of giants." First, we should leverage the pre-trained general-domain FMs by adapting them to the geography domain (Deng *et al.* 2023, Jakubik *et al.* 2023) or utilizing them as one component for the underdeveloped GeoFM. For example, geospatial tasks sometimes also require text data. Instead of training a text encoder from scratch, we can fine-tune an existing open-sourced LLM model and make it one model component of our GeoFM (Mai *et al.* 2023a). Second, we should incorporate existing, well-established GeoAI neural network modules, such as spatial representation learning modules (Mai *et al.* 2023e), satellite image encoder and decoder modules (Cong *et al.* 2022) as model components for handling respective geospatial data modalities. Third, we should reuse the established geospatial datasets as parts of the pre-training or finetuning datasets. Last but not least, when available datasets for some data modalities are sparse, we should generate some synthetic geospatial data samples by following the existing practices used in FM development if possible.

4.3. *Multimodal pre-training is the key*

One key ingredient of the success of current FMs is the large-scale self-supervised model pre-training (Brown *et al.* 2020, OpenAI 2023), especially cross-data modality pre-training (Radford *et al.* 2021, Liu *et al.* 2023a, Peng *et al.* 2023) which can enable knowledge transfer across data modalities. Since multimodality is the nature of almost all geospatial tasks, multimodal model pre-training should be the key to GeoFM development. As Mai *et al.* (2023a) pointed out, we can leverage the geospatial alignments among different data modalities (e.g., location-to-text alignment via geotagged text data, street view-to-satellite image alignment via their location metadata) as the self-supervised signals for multimodal model pre-training.

4.4. *Sustainable training framework*

It becomes a generally recognized fact that foundation models are very expensive to train and maintain and might lead to significant environmental impact (Shi *et al.* 2023). For example, the carbon cost of training a BERT language model on GPUs without hyperparameter tuning is comparable to a trans-American flight (Strubell *et al.* 2019). Analysts and technologists estimated that the critical process of training GPT-3 could cost more than $4 million[2]. And the carbon footprint of GPT-3 has been estimated to be 8.4 tons of CO_2 in a year. Touvron *et al.* (2023) reported that they used 2048 A100-80GB for approximately 5 months to develop the LLaMA model which cost 2,638 MWh and a total emission of 1,015 tCO2eq.

The high cost of FM development leads to a rather lower refreshment rate of these models which means the pre-trained foundation models can be quickly out-of-date. This is particularly challenging for GeoFMs since most geospatial tasks aim at predicting the future status of the earth given its "current" information. How to balance the high economic and environmental cost of GeoFM development and the need for a real-time updated GeoFM for multiple downstream tasks? Here, we suggest several best practices:

(1) **A sustainable evaluation framework for GeoFM**: we should develop a sustainable evaluation framework to measure the environmental impact of a given GeoFM, including factors such as energy consumption, carbon footprint, and potential contributions to geographic inequality (Shi *et al.* 2023).
(2) **Treating GeoFM as an agent but not a knowledge base**: One way to avoid a frequent model refreshment is to use GeoFM as an agent (Pesaru *et al.* 2023, Huang *et al.* 2023a, Dai *et al.* 2023, Pesaru *et al.* 2023) which tells us how to solve the given task, which tool to use, and which external knowledge bases can be used for problem-solving rather than using GeoFMs themselves as knowledge bases. Geographic knowledge that is required for a geospatial task can quickly evolve over time but the pipeline to solve a given task should remain stable. A GeoFM should be used to generate a problem-solving pipeline while we rely on a real-time updated external knowledge base such as geographic knowledge graphs (Stadler *et al.* 2012, Hoffart *et al.* 2013, Janowicz *et al.* 2022) for real-time geographic fact lookup (Liang *et al.* 2017). This can effectively lower the cost of constant GeoFM refreshment and build a sustainable training framework for GeoFM.

4.5. *GeoEthics evaluation framework*

The development of foundation models also has raised lots of ethical (Gehman *et al.* 2020, Zhao *et al.* 2018, Wang *et al.* 2023a) and privacy concerns (Rao *et al.* 2023a) from the general public. Various efforts have been contributed to quantify the toxicity (Deshpande *et al.* 2023), gender bias (Zhao *et al.* 2018), and trustworthiness (Wang *et al.* 2023a) in foundation models by developing ethics evaluation frameworks and

[2] `https://www.cnbc.com/2023/03/13/chatgpt-and-generative-ai-are-booming-but-at-a-very-expensive-price.html`

datasets. However, there are several unique GeoEthics aspects that have not been systematically studied such as geographic bias, geopolitical bias (Faisal and Anastasopoulos 2022), temporal bias (Mai *et al.* 2023a), etc. A GeoEthics evaluation framework is needed to systematically quantify these GeoEthics aspects for any given GeoFM.

5. Conclusion

In this paper, we address the "what," "why," and "how" of geo-foundation model development. As GeoFM research is in its infancy, our goal is to offer a comprehensive guideline for this promising, demanding, but also challenging field. Creating a reliable, sustainable, and ethically conscious geo-foundation model necessitates a joint endeavor among geographers, spatial data scientists, AI researchers, funding agencies, and key industry stakeholders, among others.

Acknowledgements We would like to thank Dr. Krzysztof Janowicz for his comments on the naming and definition of geo-foundation models. We would like to thank Dr. Ni Lao for his suggestion about "treating GeoFM as an agent but not a knowledge base." We would like to thank Dr. Stefano Ermon for his comments on satellite image-based vision foundation models.

References

Ayush, K., *et al.*, 2021. Geography-aware self-supervised learning. *In*: *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 10181–10190.

Balsebre, P., *et al.*, 2023. Cityfm: City foundation models to solve urban challenges. *arXiv preprint arXiv:2310.00583.*

Bommasani, R., *et al.*, 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258.*

Brown, T., *et al.*, 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877–1901.

Cai, L., *et al.*, 2023. Hyperquaternione: A hyperbolic embedding model for qualitative spatial and temporal reasoning. *GeoInformatica*, 27 (2), 159–197.

Caron, M., *et al.*, 2021. Emerging properties in self-supervised vision transformers. *In*: *Proceedings of the IEEE/CVF international conference on computer vision.* 9650–9660.

Cepeda, V.V., Nayak, G.K., and Shah, M., 2023. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *In*: *Thirty-seventh Conference on Neural Information Processing Systems.*

Chen, W., *et al.*, 2013. A synergistic framework for geographic question answering. *In*: *2013 IEEE seventh international conference on semantic computing.* IEEE, 94–99.

Cole, E., *et al.*, 2023. Spatial implicit neural representations for global-scale species mapping. *In*: *International Conference on Machine Learning.* PMLR.

Cong, Y., *et al.*, 2022. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35, 197–211.

Dai, H., *et al.*, 2023. Ad-autogpt: An autonomous gpt for alzheimer's disease infodemiology. *arXiv preprint arXiv:2306.10095.*

Deng, C., *et al.*, 2023. Learning a foundation language model for geoscience knowledge understanding and utilization. *arXiv preprint arXiv:2306.05064.*

Deng, J., *et al.*, 2009. Imagenet: A large-scale hierarchical image database. *In*: *2009 IEEE conference on computer vision and pattern recognition.* Ieee, 248–255.

Deshpande, A., *et al.*, 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*.

Dosovitskiy, A., *et al.*, 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Elmustafa, A., *et al.*, 2022. Understanding economic development in rural africa using satellite imagery, building footprints and deep models. *In*: *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*. 1–4.

Faisal, F. and Anastasopoulos, A., 2022. Geographic and geopolitical biases of language models. *arXiv preprint arXiv:2212.10408*.

Fuller, A., Millard, K., and Green, J.R., 2023. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. *In*: *Thirty-seventh Conference on Neural Information Processing Systems*.

Gehman, S., *et al.*, 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv e-prints*, arXiv–2009.

Goodchild, M.F. and Li, W., 2021. Replication across space and time must be weak in the social and environmental sciences. *Proceedings of the National Academy of Sciences*, 118 (35), e2015759118.

Ha, D. and Eck, D., 2018. A neural representation of sketch drawings. *In*: *International Conference on Learning Representations*.

He, K., *et al.*, 2016. Deep residual learning for image recognition. *In*: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

He, Y., *et al.*, 2021. Spatial-temporal super-resolution of satellite imagery via conditional pixel synthesis. *Advances in Neural Information Processing Systems*, 34, 27903–27915.

Helber, P., *et al.*, 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12 (7), 2217–2226.

Henry Wai-Chung, Y., 2001. Redressing the geographical bias in social science knowledge.

Hoffart, J., *et al.*, 2013. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artificial intelligence*, 194, 28–61.

Horawalavithana, S., *et al.*, 2022. Foundation models of scientific knowledge for chemistry: Opportunities, challenges and lessons learned. *In*: *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*. 160–172.

Hu, E.J., *et al.*, 2021. Lora: Low-rank adaptation of large language models. *In*: *International Conference on Learning Representations*.

Hu, X., *et al.*, 2023a. Location reference recognition from texts: A survey and comparison. *ACM Computing Surveys*, 56 (5), 1–37.

Hu, Y., *et al.*, 2023b. Geo-knowledge-guided gpt models improve the extraction of location descriptions from disaster-related social media messages. *International Journal of Geographical Information Science*, 1–30.

Huang, Q., *et al.*, 2023a. Benchmarking large language models as ai research agents. *In*: *NeurIPS 2023 Foundation Models for Decision Making Workshop*.

Huang, S., *et al.*, 2023b. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*.

Huang, Z., *et al.*, 2019. Geosqa: A benchmark for scenario-based question answering in the geography domain at high school level. *In*: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5866–5871.

Jakubik, J., *et al.*, 2023. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*.

Janowicz, K., *et al.*, 2020. Geoai: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond.

Janowicz, K., *et al.*, 2022. Know, know where, knowwheregraph: A densely connected, cross-domain knowledge graph and geo-enrichment service stack for applications in environmental intelligence. *AI Magazine*, 43 (1), 30–39.

Jean, N., *et al.*, 2019. Tile2vec: Unsupervised representation learning for spatially distributed data. *In*: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 33, 3967–3974.

Jiang, C., *et al.*, 2019. DDSL: Deep differentiable simplex layer for learning geometric signals. *In*: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8769–8778.

Kang, Y., *et al.*, 2021. Understanding house price appreciation using multi-source big geo-data and machine learning. *Land Use Policy*, 111, 104919.

Kenton, J.D.M.W.C. and Toutanova, L.K., 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *In*: *Proceedings of NAACL-HLT*. 4171–4186.

Kirillov, A., *et al.*, 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Klemmer, K., *et al.*, 2023. Towards global, general-purpose pretrained geographic location encoders. *In*: *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*.

Kuhn, W., 2012. Core concepts of spatial information for transdisciplinary research. *International Journal of Geographical Information Science*, 26 (12), 2267–2276.

Lacoste, A., *et al.*, 2023. Geo-bench: Toward foundation models for earth monitoring.

Lam, D., *et al.*, 2018. xview: Objects in context in overhead imagery. *arXiv preprint arXiv:1802.07856*.

Latif, E., *et al.*, 2023. Artificial general intelligence (agi) for education. *arXiv preprint arXiv:2304.12479*.

Lee, G.G., *et al.*, 2023. Multimodality of ai for education: Towards artificial general intelligence. *arXiv preprint arXiv:2312.06037*.

Lee, J., *et al.*, 2021. Predicting livelihood indicators from crowdsourced street level images. *In*: *Proceedings of the AAAI Conference on Artificial Intelligence*.

Lewis, M., *et al.*, 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Li, C., *et al.*, 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *In*: *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems*.

Li, W., Hsu, C.Y., and Hu, M., 2021. Tobler's first law in geoai: A spatially explicit deep learning model for terrain feature detection under weak supervision. *Annals of the American Association of Geographers*, 111 (7), 1887–1905.

Li, Z., *et al.*, 2023b. Geolm: Empowering language models for geospatially grounded language understanding. *In*: *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Liang, C., *et al.*, 2017. Neural symbolic machines: Learning semantic parsers on freebase with weak supervision. *In*: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 23–33.

Lin, T.Y., *et al.*, 2014. Microsoft coco: Common objects in context. *In*: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

Liu, H., *et al.*, 2023a. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Liu, Z., *et al.*, 2023b. Transformation vs tradition: Artificial general intelligence (agi) for arts and humanities. *arXiv preprint arXiv:2310.19626*.

Liu, Z., *et al.*, 2022. Geoparsing: Solved or biased? an evaluation of geographic biases in geoparsing. *AGILE: GIScience Series*, 3, 9.

Lobry, S., Demir, B., and Tuia, D., 2021. Rsvqa meets bigearthnet: a new, large-scale, visual question answering dataset for remote sensing. *In*: *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE, 1218–1221.

Lobry, S., *et al.*, 2020. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58 (12), 8555–8566.

Lu, G., *et al.*, 2023. Agi for agriculture. *arXiv preprint arXiv:2304.06136*.

Ma, J. and Wang, B., 2023. Towards foundation models of biological image segmentation. *Nature Methods*, 20 (7), 953–955.

Mac Aodha, O., Cole, E., and Perona, P., 2019. Presence-only geographical priors for fine-grained image classification. *In*: *Proceedings of the IEEE International Conference on Computer Vision*. 9596–9606.

Mai, G., *et al.*, 2022a. Towards a foundation model for geospatial artificial intelligence (vision paper). *In*: *Proceedings of the 30th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 1–4.

Mai, G., *et al.*, 2022b. Symbolic and subsymbolic geoai: Geospatial knowledge graphs and spatially explicit machine learning. *Transactions in GIS*, 26 (8), 3118–3124.

Mai, G., *et al.*, 2023a. On the opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv preprint arXiv:2304.06798*.

Mai, G., *et al.*, 2020a. Se-kge: A location-aware knowledge graph embedding model for geographic question answering and spatial semantic lifting. *Transactions in GIS*, 24 (3), 623–655.

Mai, G., *et al.*, 2018. Poireviewqa: A semantically enriched poi retrieval and question answering dataset. *In*: *Proceedings of the 12th Workshop on Geographic Information Retrieval*. 1–2.

Mai, G., *et al.*, 2022c. A review of location encoding for geoai: methods and applications. *International Journal of Geographical Information Science*, 36 (4), 639–673.

Mai, G., *et al.*, 2020b. Multi-scale representation learning for spatial feature distributions using grid cells. *In*: *International Conference on Learning Representations*.

Mai, G., *et al.*, 2021. Geographic question answering: Challenges, uniqueness, classification, and future directions. *AGILE: GIScience series*, 2, 8.

Mai, G., *et al.*, 2023b. Towards general-purpose representation learning of polygonal geometries. *GeoInformatica*, 27 (2), 289–340.

Mai, G., *et al.*, 2023c. Csp: Self-supervised contrastive spatial pre-training for geospatial-visual representations. *In*: *the Fortieth International Conference on Machine Learning (ICML 2023)*.

Mai, G., *et al.*, 2023d. Ssif: Learning continuous image representation for spatial-spectral super-resolution. *arXiv preprint arXiv:2310.00413*.

Mai, G., Li, Z., and Lao, N., 2023e. Spatial representation learning in geoai. *In*: *Handbook of geospatial artificial intelligence*. CRC Press, 99–120.

Mai, G., *et al.*, 2023f. Sphere2vec: A general-purpose location representation learning over a spherical surface for large-scale geospatial predictions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202, 439–462.

Mai, G., *et al.*, 2020c. Relaxing unanswerable geographic questions using a spatially explicit knowledge graph embedding model. *In*: *Geospatial Technologies for Local and Regional Development: Proceedings of the 22nd AGILE Conference on Geographic Information Science 22*. Springer, 21–39.

Manas, O., *et al.*, 2021. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. *In*: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9414–9423.

Manvi, R., *et al.*, 2024. Geollm: Extracting geospatial knowledge from large language models. *In*: *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*.

Mei, S., *et al.*, 2020. Spatial and spectral joint super-resolution using convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 58 (7), 4590–4603.

Mishra, A., Mishra, N., and Agrawal, A., 2010. Context-aware restricted geographical domain question answering system. *In*: *2010 international conference on computational intelligence and communication networks*. IEEE, 548–553.

Moor, M., *et al.*, 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616 (7956), 259–265.

Müller, M., *et al.*, 2020. Super-resolution of multispectral satellite images using convolutional neural networks. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 1, 33–40.

Nguyen, T., *et al.*, 2023. Climax: A foundation model for weather and climate. *In*: *the Fortieth*

International Conference on Machine Learning (ICML 2023).
Oord, A.v.d., Li, Y., and Vinyals, O., 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748.*
OpenAI, 2022. Introducing ChatGPT — openai.com. `https://openai.com/blog/chatgpt`. [Accessed 30-11-2022].
OpenAI, 2023. Gpt-4 technical report. *arXiv*, 2303–08774.
Ouyang, L., *et al.*, 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
Peng, Z., *et al.*, 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824.*
Pesaru, A., Gill, T.S., and Tangella, A.R., 2023. Ai assistant for document management using lang chain and pinecone. *International Research Journal of Modernization in Engineering Technology and Science.*
Radford, A., *et al.*, 2021. Learning transferable visual models from natural language supervision. *In*: *International conference on machine learning.* PMLR, 8748–8763.
Radford, A., *et al.*, 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1 (8), 9.
Ramesh, A., *et al.*, 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125.*
Randell, D.A., Cui, Z., and Cohn, A.G., 1992. A spatial logic based on regions and connection. *KR*, 92, 165–176.
Rao, J., *et al.*, 2020. LSTM-TrajGAN: A deep learning approach to trajectory privacy protection. *In*: *GIScience 2020.* 12:1–12:17.
Rao, J., *et al.*, 2023a. Building privacy-preserving and secure geospatial artificial intelligence foundation models. *In*: *Proceedings of the 31th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems.*
Rao, J., Gao, S., and Zhu, S., 2023b. Cats: Conditional adversarial trajectory synthesis for privacy-preserving trajectory data publication using deep learning approaches. *International Journal of Geographical Information Science*, 1–37.
Regalia, B., Janowicz, K., and McKenzie, G., 2019. Computing and querying strict, approximate, and metrically refined topological relations in linked geographic data. *Transactions in GIS*, 23 (3), 601–619.
Rempe, D., *et al.*, 2023. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. *In*: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 13756–13766.
Renz, J. and Nebel, B., 2007. Qualitative spatial reasoning using constraint calculi. *In*: *Handbook of spatial logics.* Springer, 161–215.
Roberts, J., *et al.*, 2023. Gpt4geo: How a language model sees the world's geography. *arXiv preprint arXiv:2306.00020.*
Rombach, R., *et al.*, 2022. High-resolution image synthesis with latent diffusion models. *In*: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 10684–10695.
Schaefer, F.K., 1953. Exceptionalism in geography: A methodological examination. *Annals of the Association of American geographers*, 43 (3), 226–249.
Scheider, S., *et al.*, 2021. Geo-analytical question-answering with gis. *International Journal of Digital Earth*, 14 (1), 1–14.
Shi, M., *et al.*, 2023. Thinking geographically about ai sustainability. *AGILE: GIScience Series*, 4, 42.
Soni, A. and Boddhu, S., 2022. Finding map feature correspondences in heterogeneous geospatial datasets. *In*: *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Geospatial Knowledge Graphs.* 7–16.
Stadler, C., *et al.*, 2012. Linkedgeodata: A core for a web of spatial open data. *Semantic Web*, 3 (4), 333–354.
Strubell, E., Ganesh, A., and McCallum, A., 2019. Energy and policy considerations for deep

learning in nlp. *arXiv preprint arXiv:1906.02243*.

Sumbul, G., *et al.*, 2019. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. *In*: *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 5901–5904.

Tan, C., *et al.*, 2023. On the promises and challenges of multimodal foundation models for geographical, environmental, agricultural, and urban planning applications. *arXiv preprint arXiv:2312.17016*.

Team, G., *et al.*, 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Touvron, H., *et al.*, 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vaswani, A., *et al.*, 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Veer, R.v., Bloem, P., and Folmer, E., 2018. Deep learning for classification tasks on geospatial vector polygons. *arXiv preprint arXiv:1806.03857*.

Wang, B., *et al.*, 2023a. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *In*: *Thirty-seventh Conference on Neural Information Processing Systems*.

Wang, J., Hu, Y., and Joseph, K., 2020. Neurotpr: A neuro-net toponym recognition model for extracting locations from social media messages. *Transactions in GIS*, 24 (3), 719–735.

Wang, X., *et al.*, 2023b. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*.

Wedi, N., *et al.*, 2015. *The modelling infrastructure of the integrated forecasting system: Recent advances and future challenges*. European Centre for Medium-Range Weather Forecasts.

Wei, J., *et al.*, 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Weinstein, B.G., *et al.*, 2021. A remote sensing derived data set of 100 million individual tree crowns for the national ecological observatory network. *Elife*, 10, e62922.

Xie, Y., *et al.*, 2023. Geo-foundation models: Reality, gaps and opportunities (vision paper). *In*: *Proceedings of the 31th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*.

Yan, X., *et al.*, 2021. Graph convolutional autoencoder model for the shape coding and cognition of buildings in maps. *International Journal of Geographical Information Science*, 35 (3), 490–512.

Yan, X., *et al.*, 2019. A graph convolutional neural network for classification of building patterns using spatial vector data. *ISPRS journal of photogrammetry and remote sensing*, 150, 259–273.

Yang, Y. and Newsam, S., 2010. Bag-of-visual-words and spatial extensions for land-use classification. *In*: *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*. 270–279.

Yeh, C., *et al.*, 2020. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications*, 11 (1), 2583.

Yu, W. and Chen, Y., 2022. Filling gaps of cartographic polylines by using an encoder–decoder model. *International Journal of Geographical Information Science*, 1–26.

Yue, Y., *et al.*, 2023. Connecting the Dots: Floorplan Reconstruction Using Two-Level Queries. *In*: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, F., *et al.*, 2019. Social sensing from street-level imagery: A case study in learning spatio-temporal urban mobility patterns. *ISPRS journal of photogrammetry and remote sensing*, 153, 48–58.

Zhang, F., *et al.*, 2018. Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning*, 180, 148–160.

Zhang, J., *et al.*, 2023a. Text2seg: Remote sensing image semantic segmentation via text-guided visual foundation models. *arXiv preprint arXiv:2304.10597*.

Zhang, S., *et al.*, 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhang, Y., *et al.*, 2023b. Geogpt: Understanding and processing geospatial tasks through an autonomous gpt. *arXiv preprint arXiv:2307.07930*.

Zhao, J., *et al.*, 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

Zhu, R., *et al.*, 2022. Reasoning over higher-order qualitative spatial relations via spatially explicit neural networks. *International Journal of Geographical Information Science*, 36 (11), 2194–2225.