Large Language Models as a New Modality for Generalizable Earth Data Monitoring

Climate Change AI

Tong Nie, Junlin He, and Wei Ma tong.nie@connect.polyu.hk

ICLR

Introduction

SUSTAINABLE DEVELOPMENT GOALS

Background:

- Earth observation data are critical for monitoring progress toward **Sustainable Development Goals (SDGs)**.
- Persistent challenges in **accessibility (equity)**, integration of **multimodal** data, and **geographic bias** hinder comprehensive **global assessments**.

Motivation

Research Gaps:

- Traditional ground-based surveying systems exhibit significant **geographic disparities**, with global south countries lacking adequate infrastructure.
- Satellite image with machine learning (SIML) approaches are cost-effective, but they still face **challenges:**
- High-resolution satellite images are usually private with high technical barriers to access **(inequity)**; Unbalanced data distribution **(geographic bias)**.
- Remote sensing features only observe the earth's **"physical appearance"**, hard to reflect socioeconomics and population dynamics.
- **A new globally accessible data sources is needed to measure SDGs.**

Our Solution

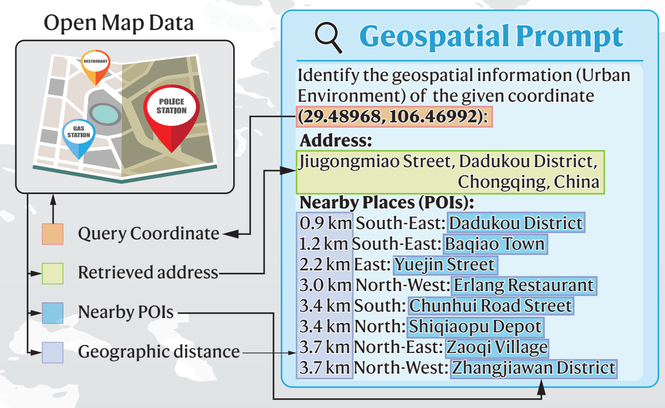
Efficient, generalizable, and accessible

Large Language Models as a New Data Modality:

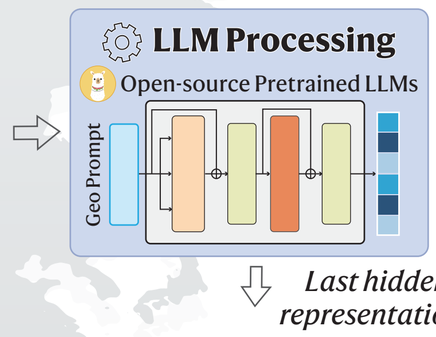
- By training on data from the entire internet, LLMs acquire **extensive geographical knowledge**, provide insights into socio-economic development.
- We develop an approach to **extract geographic knowledge from LLMs as vectors**, serving as a **new data modality to complement SIML**.
- The representation for each location is applied to estimate a variety of **earth monitoring** indicators with a **simple regression**.

Methodology

Stage 1: generating geolocation prompts for coordinates from open map data



Stage 2: generating embeddings for geo-prompt from pre-trained LLMs

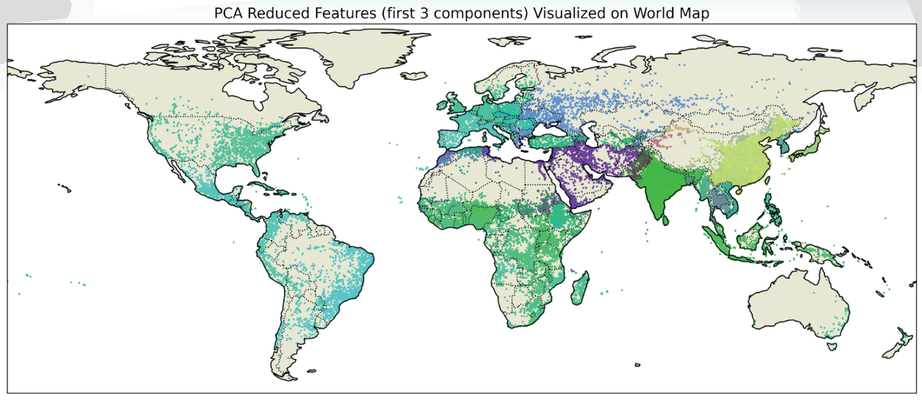


Stage 3: using simple regressors to estimate earth observation variables

$$y_i^{(t)} = \mathbf{z}_i \beta^{(t)} + \epsilon_i^{(t)}, \ \forall P_i \in P_{\text{train}},$$

$$\min_{\beta^{(t)}} \|y_i^{(t)} - \mathbf{z}_i \beta^{(t)}\|_2^2 + \alpha^{(t)} \|\beta^{(t)}\|_2^2,$$

$$t = \{1, \ldots, T\} \quad \text{for each task}$$



PCA Structure

Meaningful geospatial representation on the earth

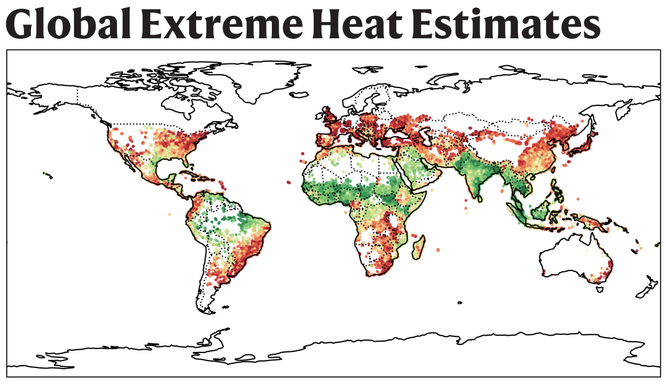
Experiments & Findings

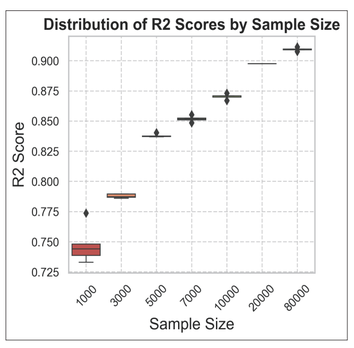
SDGs Related Earth Data Monitoring (Estimation)

13 CLIMATE ACTION

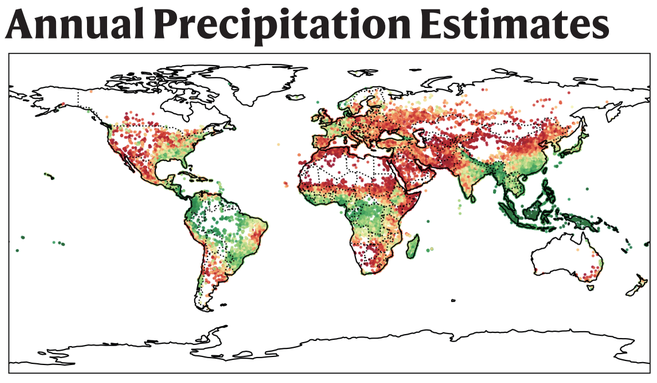
15 LIFE ON LAND

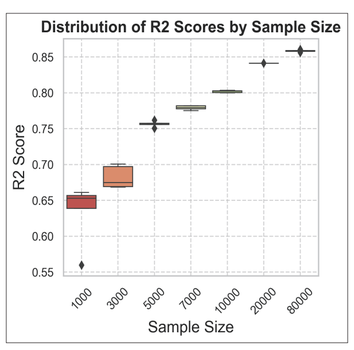
Global Extreme Heat Estimates



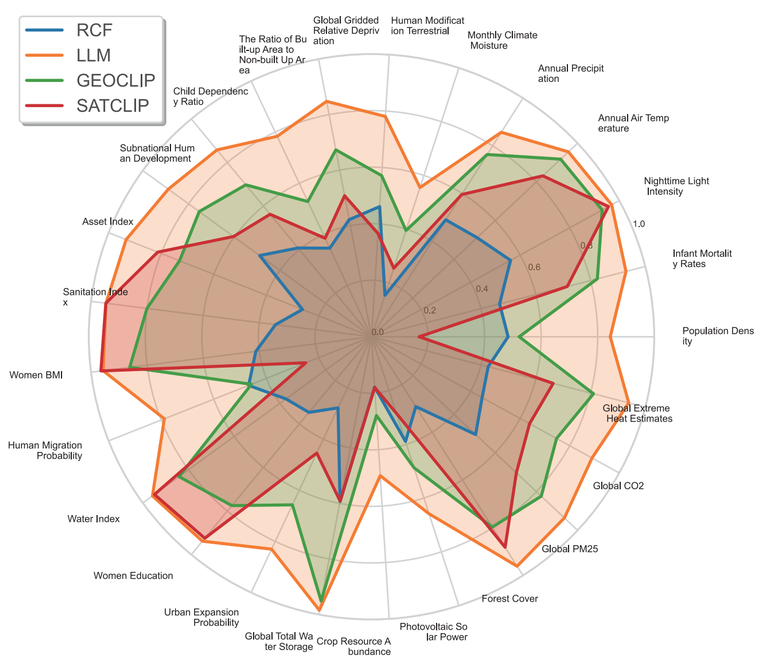


Annual Precipitation Estimates





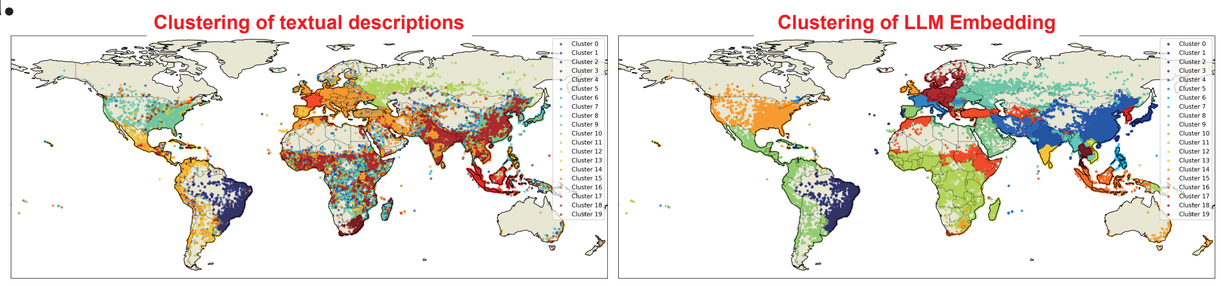
Comparison of Methods Across 25 Tasks (Estimation R^2)



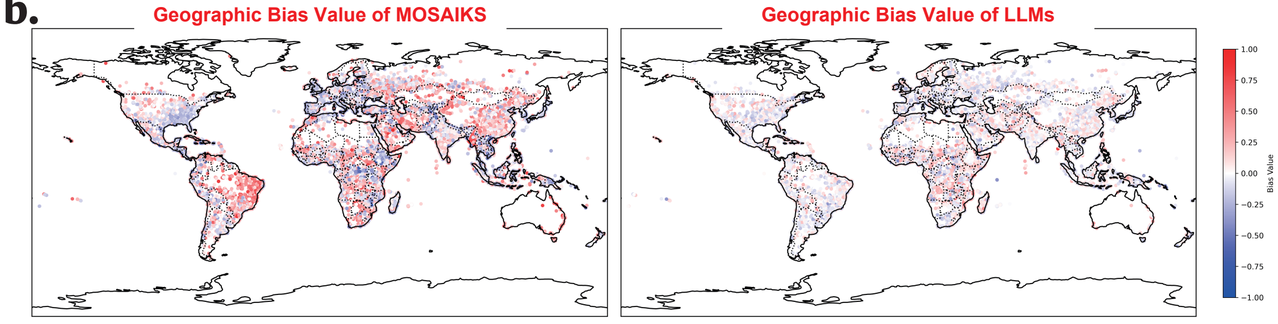
Takeaways

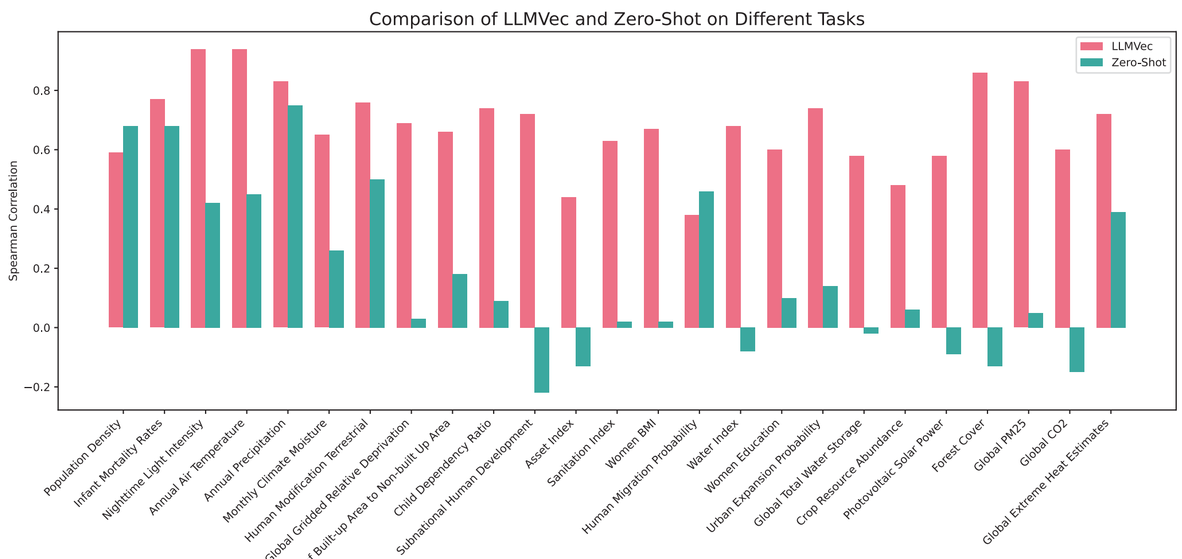
- Geographical representations of LLMs can predict earth observation variables closely tied to SDGs at the global scale with linear regressor.
- It outperforms SIML and general pretrained earth models on 25 SDGs-related tasks, without any fine-tuning of the backbone LLMs.
- This embedding is geospatially aware, preserving rich spatial semantics in feature space, and exhibits less geographical bias compared to satellite image based features.

a.



b.





SCAN TO VIEW MORE OF OUR WORK

THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學

TONGJI UNIVERSITY 同濟大學

TOPS

Mobility AI Lab

WE ARE INTERESTED IN:

- AI for Social Good
- Urban Science
- Smart Transportation
- Large Language Models

wei.w.ma@polyu.edu.hk