GeoGrid-Bench: Can Foundation Models Understand Multimodal Gridded Geo-Spatial Data?

Bowen Jiang[1, 2] **Yangxinyu Xie**[1, 2] **Xiaomeng Wang**[1] **Jiashu He**[1]
John K. Hutchison[2] **Camillo J. Taylor**[1] **Tanwi Mallick**[2]
University of Pennsylvania[1] Argonne National Laboratory[2]
Philadelphia, PA, 19104 Lemont, IL, 60439
{bwjiang@seas, xinyux@wharton, cjtaylor@seas}.upenn.edu, tmallick@anl.gov

Abstract

We present GeoGrid-Bench, a benchmark designed to evaluate the ability of foundation models to understand geo-spatial data in the grid structure. Geo-spatial datasets pose distinct challenges due to their dense numerical values, strong spatial and temporal dependencies, and unique multimodal representations including tabular data, heatmaps, and geographic visualizations. To assess how foundation models can support scientific research in this domain, GeoGrid-Bench features large-scale, real-world data covering 16 climate variables across 150 locations and extended time frames. The benchmark includes approximately 27,000 question-answer pairs, systematically generated from 8 domain expert-curated templates to reflect practical tasks encountered by human scientists. These range from basic queries at a single location and time to complex spatiotemporal comparisons across regions and periods. Our evaluation reveals that vision-language models perform best overall, and we provide a fine-grained analysis of the strengths and limitations of different foundation models in different geo-spatial tasks. This benchmark offers clearer insights into how foundation models can be effectively applied to geo-spatial data analysis and used to support scientific research.[1]

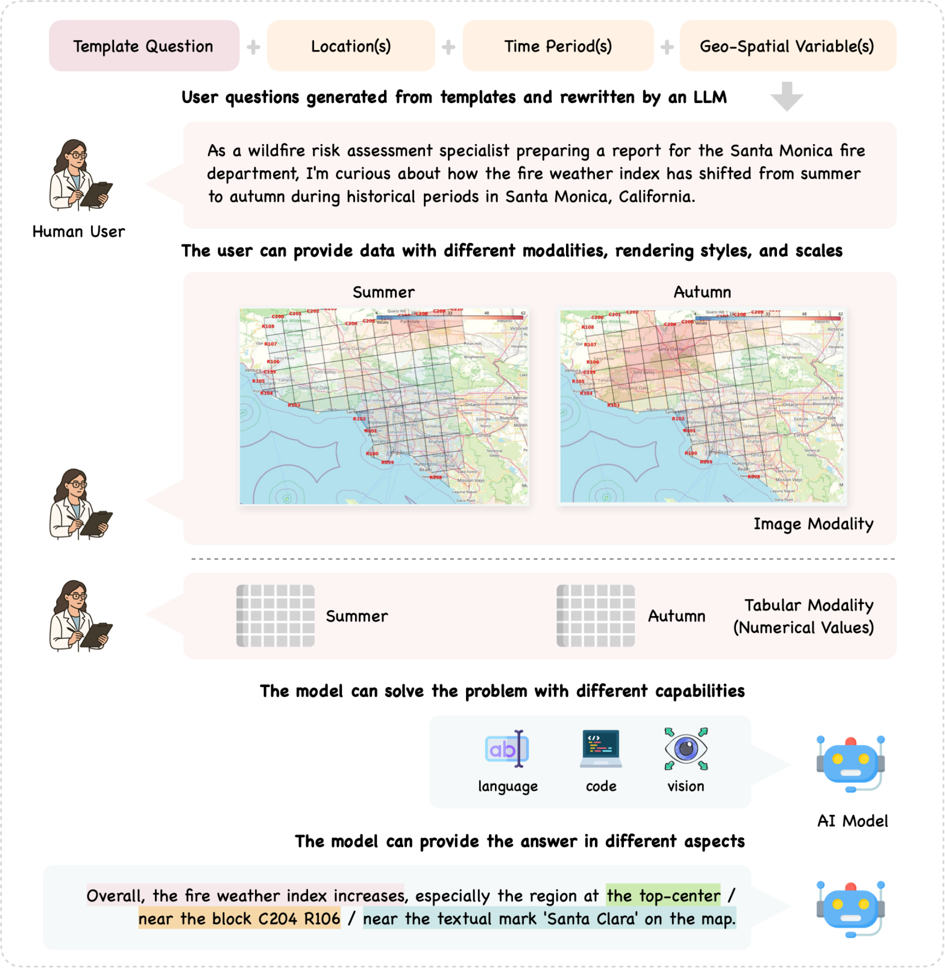
1 Introduction

Foundation models have demonstrated transformative capabilities across diverse domains, ranging from language and vision to programming and reasoning (Hurst et al., 2024; Jaech et al., 2024; Jiang et al., 2024b,c,a; Balachandran et al., 2024; Jiang et al., 2024d; He et al., 2024a). Their rapid advancement has naturally inspired research exploring their utility in scientific contexts, particularly in critical fields like climate science and natural hazard assessment (Mai et al., 2022; Xie et al., 2024, 2025; Nguyen et al., 2023; Mai et al., 2023; de Rijke et al., 2025; Mallick et al., 2025), where accurate, data-intensive decision-making can profoundly impact human well-being.

Geo-spatial data pose distinct challenges for foundation models due to their inherent spatio-temporal dependencies and exceptionally high data density. Unlike typical tabular records for knowledge retrieval (Zhang et al., 2023a; Pasupat & Liang, 2015; Zhang et al., 2025) or natural images, climate data exists in structured, gridded formats with complex, interconnected numerical values often represented through modalities such as tables, heatmaps, or geographic images spanning across space and time. These data are typically organized in highly structured, gridded formats that encode interconnected numerical values across spatial and temporal dimensions. Each data point is not an isolated unit but part of a dense, multi-dimensional array that reflects physical processes,

[1] All code and data are publicly available at our Github repository https://github.com/bowen-upenn/GeoGrid_Bench and Huggingface https://huggingface.co/datasets/bowen-upenn/GeoGrid_Bench.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.


Template Question
+
Location(s)
+
Time Period(s)
+
Geo-Spatial Variable(s)
User questions generated from templates and rewritten by an LLM
Human User
The user can provide data with different modalities, rendering styles, and scales
Summer
Autumn
Image Modality
Summer
Autumn
Tabular Modality
(Numerical Values)
The model can solve the problem with different capabilities
ab
language
code
vision
AI Model
The model can provide the answer in different aspects

As a wildfire risk assessment specialist preparing a report for the Santa Monica fire department, I'm curious about how the fire weather index has shifted from summer to autumn during historical periods in Santa Monica, California.

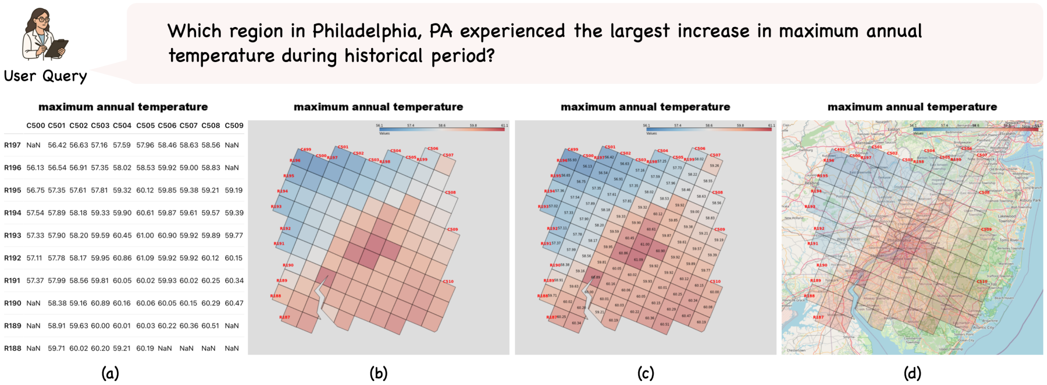
Overall, the fire weather index increases, especially the region at the top-center / near the block C204 R106 / near the textual mark 'Santa Clara' on the map.

Figure 1: **Overview of GeoGrid-Bench.** The benchmark features questions generated from templates that vary by location, time period, and climate variable, then rewritten with natural language context. Each question is paired with multimodal input—either heatmaps as images or tabular grids of numerical values. We evaluate models on their ability to solve the queries through different modalities—natural language, code, or vision. Ground-truth answers capture find-grained aspects like overall trends, spatial references (from top-left to lower-right), coordinate references (row and column indices), and label references (textual marks on the maps), whenever available.

environmental interactions, or geographical phenomena evolving over time. Meanwhile, models can also easily get lost in the context (Liu et al., 2023) with overwhelming volumes of values per sample.

Informed decision-making in fields such as disaster response, climate science, and urban development depends on the ability to detect and interpret patterns across regions and over time. However, there remains a lack of benchmarks that directly address the unique challenges posed by geo-spatial gridded data. Most existing efforts docus on object detection, semantic segmentation, object counting, captioning, or scene understanding of Earth observation images (Lacoste et al., 2023; Danish et al., 2024; Zhang & Wang, 2024; Zheng et al., 2023; Wang et al., 2024a; Muhtar et al., 2024; Bazi et al., 2024; Kuckreja et al., 2024), function calls to the Geographic Information System (GIS) or SQL queries for data retrieval (Krechetova & Kochedykov, 2025; Jiang & Yang, 2024; Ning et al., 2025; Mooney et al., 2023; Zhang et al., 2023b), or simplified query setups that overlook the spatial-temporal complexities in practical geo-spatial analysis (Bhandari et al., 2023).

To understand how foundation models can assist geo-spatial data analysis, we introduce GeoGrid-Bench, a benchmark explicitly designed to evaluate model performance on multimodal, real-world geo-spatial data. We adopt domain expert-curated query templates to reflect realistic questions that


User Query
maximum annual temperature
maximum annual temperature
maximum annual temperature
maximum annual temperature
(a)
(b)
(c)
(d)

Which region in Philadelphia, PA experienced the largest increase in maximum annual temperature during historical period?

	C500	C501	C502	C503	C504	C505	C506	C507	C508	C509
R197	NaN	56.42	56.63	57.16	57.59	57.96	58.46	58.63	58.56	NaN
R196	56.13	56.54	56.91	57.35	58.02	58.53	59.92	59.00	58.83	NaN
R195	56.75	57.35	57.61	57.81	59.32	60.12	59.85	59.38	59.21	59.19
R194	57.54	57.89	58.18	59.33	59.90	60.61	59.87	59.61	59.57	59.39
R193	57.33	57.90	58.20	59.59	60.45	61.00	60.90	59.92	59.89	59.77
R192	57.11	57.78	58.17	59.95	60.86	61.09	59.92	59.92	60.12	60.15
R191	57.37	57.99	58.56	59.81	60.05	60.02	59.93	60.02	60.25	60.34
R190	NaN	58.38	59.16	60.89	60.16	60.06	60.05	60.15	60.29	60.47
R189	NaN	58.91	59.63	60.00	60.01	60.03	60.22	60.36	60.51	NaN
R188	NaN	59.71	60.02	60.20	59.21	60.19	NaN	NaN	NaN	NaN

Figure 2: We prepare every data sample in one of the four formats: (a) 2D table as a textual string. (b) standalone heatmap; (c) heatmap with overlaid numerical annotations at each grid cell; (d) heatmap overlaid on an actual geographic base map. These formats reflect real-world climate data practices and differ markedly from typical natural images seen by foundation models. More in Appendix B.

practitioners would encounter in geo-spatial analysis—providing data in both tabular and image formats. These tasks range from simple queries about a fixed location and time to more complex analyses involving multiple locations and temporal comparisons. For each template, we develop oracle code that is applied uniformly to all query instances, enabling scalable and consistent generation of question-answer pairs. Our contributions can be summarized as follows:

Large-scale, real-world data: A domain-centric benchmark built on large-scale, real-world climate projection data, presented in multimodal formats commonly used by actual practitioners, including structured numerical tables and geographic visualizations.

Scalable query generation: A systematic user query generation pipeline based on domain expert-designed templates, reflecting diverse and realistic scientific challenges.

Comprehensive evaluation: Evaluation of foundation models with language, coding, multimodal, and reasoning capabilities across find-grained answer aspects and data modalities to diagnose their strengths and weaknesses in geo-spatial analysis tasks.

Through comprehensive evaluations, we find that visualizing dense, gridded geo-spatial data as heatmaps is the most accessible format for existing foundation models to interpret. In contrast, models struggle to generate flawless code for completing these tasks. Across all model types, identifying broad trends proves easier than making fine-grained regional distinctions, and models exhibit varying strengths and weaknesses depending on the task. With GeoGrid-Bench, we aim to shed light on the strengths and limitations of current foundation models when applied to multimodal geo-spatial data, a core yet underexplored format in climate science. Our goal is to support and advance the development of practical AI-assisted tools that can aid scientific research and decision-making.

2 GeoGrid-Bench: Overview of Data Features and Tasks

GeoGrid-Bench aims to reflect the real-world challenges that scientists face when analyzing geo-spatial data at scale. To achieve this, it features *large-scale, real-world* geo-spatial data sourced and sampled from ClimRR (Argonne National Laboratory, 2023), capturing the complexity of environmental conditions across 150 locations in North America. ClimRR has demonstrated practical utility across multiple sectors, supporting hazard mitigation planning in Kentucky, climate risk assessments by utility companies, and infrastructure planning by engineering firms (TechBrew, 2025; Center for Climate and Energy Solutions, 2025), while its high-resolution data powers decision support tools like the Geospatial Energy Mapper (Argonne National Laboratory, 2025) for national-scale energy and resilience planning. The grid spans 16 diverse climate variables, such as temperature extremes, precipitation, wind speeds, humidity, fire weather indices, and degree days. An overview of user-model interaction is shown in Figure 1.

GeoGrid-Bench is built to capture the unique grid structure. Climate projection data are typically organized across spatial grids and time sequences, resulting in dense, high-dimensional arrays. The data is inherently interconnected, with each point influenced by its geographic neighbors and historical context. This structure poses unique challenges: models must capture spatio-temporal dependencies and handle variability across scales to derive meaningful insights.

Geo-spatial data is also inherently multimodal, presented as tabular data, heatmaps, or geographic visualizations, with each format sharing alignment across a spatial grid structure. Each grid cell encodes a rich array of numerical data that captures localized atmospheric behavior and climate dynamics over time. This multimodal grid structure makes our GeoGrid-Bench an ideal testbed for foundation models designed to reason across space, time, and modality. To perform well, foundation models must integrate spatial context from neighboring cells, understand temporal trends across multi-year projections, and interpret information presented in diverse formats and patterns. GeoGrid-Bench reflects this complexity and we show examples of the data formats in Figure 2.

Templates that require one data frame
1. Which region in the {location1} experienced the largest increase in {variable1} during {time_frame1}?
Templates that require two data frames
2. How has {variable1} changed between {time_frame1} and {time_frame2} in the {location1}?
3. What is the correlation between {variable1} and {variable2} in the {location1} during {time_frame1}?
4. How does {variable1} compare between {location1} and {location2} during {time_frame1}?
Templates that require four data frames
5. What is the *seasonal* variation of {climate_variable1} in {location1} during {time_frame1}?
6. Which *season* in {time_frame1} saw the highest levels of {variable1} in {location1}?
7. Which of {location1} or {location2} experienced a greater change in {variable1} throughout {time_frame1} and {time_frame2}?
Templates that require eight data frames
8. How does the *seasonal* variation of {variable1} in {location1} compare to that in {location2} for {time_frame1}?

Table 1: **Template questions in GeoGrid-Bench.** We develop those questions with domain experts. Each question includes placeholders for one or two locations, time frames, and geo-spatial variables. This design enables scalable question construction while capturing varying levels of complexity based on the number of data frames involved.

To capture the wide range of questions concerning practitioners at the forefront of geo-spatial analysis, we surveyed 13 domain experts in natural hazard risk domains, resulting in 8 template questions based on their input (Table 1) and around 27,000 query instances in GeoGrid-Bench. Each template includes placeholders based at one or two geographic locations, time frames, and climate variables, requiring one to eight data frames. This design allows us to generate a scalable set of scientifically concrete queries that reflect analytical goals. Specifically, GeoGrid-Bench evaluates the following capabilities of foundation models: (1) **Identifying regions with the most significant patterns.** This is crucial for disaster response and monitoring, helping detect hotspots that need timely action. (2) **Comparing data across different locations and times.** This is essential for uncovering spatial disparities, understanding regional dynamics, and tracking changes over time. (3) **Analyzing temporal trends and seasonal variations.** This is essential for practitioners to anticipate recurring patterns and detect long-term changes to make informed decisions. (4) **Interpreting data in multimodal formats.** This is essential for understanding the ability of foundation models to interpret real-world geo-spatial data that is multimodal in nature.

Full List of Climate Variables in GeoGrid-Bench

Maximum Annual Temperature, Minimum Annual Temperature, Consecutive Days with No Precipitation, Cooling Degree Days, Fire Weather Index, Maximum Daily Heat Index, Maximum Seasonal Heat Index, Number of Days with Daily Heat Index > 95°F/105°F/115°F/125°F, Heating Degree, Annual Total Precipitation, Maximum Seasonal Temperature, Minimum Seasonal Temperature, Wind Speed.

3 Constructing GeoGrid-Bench At Scale

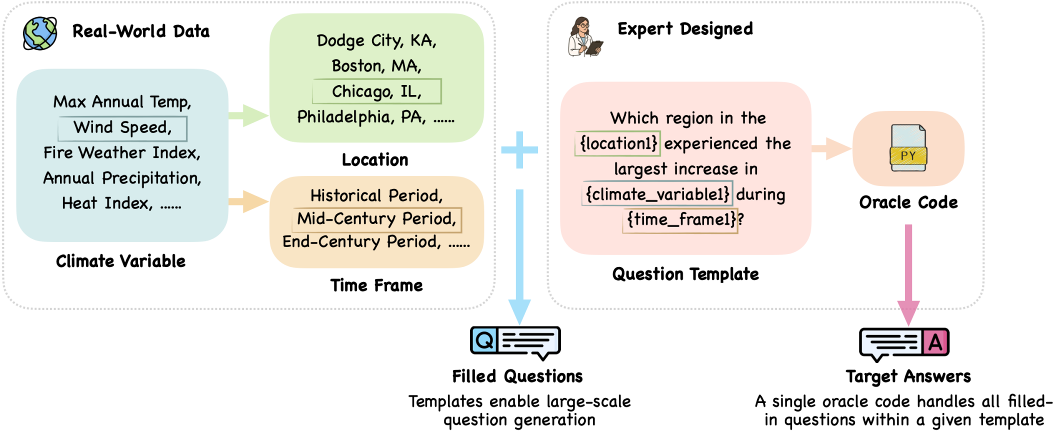


Figure 3: **Overview of the example curation process.** Each example in GeoGrid-Bench is constructed by combining a query template with sampled climate variables, locations, and time frames from real-world climate data. Each template is paired with a corresponding oracle code that deterministically generates target answers for all filled-in question instances under that template.

GeoGrid-Bench features diverse real-world geo-spatial data We illustrate our sample curation process in Figure 3. Each data sample is formed by extracting a specific climate-location-time slice from the ClimRR dataset. We sample from the 16 climate variables listed above. For each climate variable, we select around 50 locations where this climate variable is the most prominent, resulting in a total of 150 distinct locations across all climate variables, a subset of ClimRR. For example, the benchmark includes more regions in Southern California for wildfire risk, while precipitation-related examples are more concentrated in the Pacific Northwest to reflect region-specific climate concerns.

We render each data sample in either a **tabular** or **image** format, both structured over a spatial grid. For a given location and its longitude and latitude, we retrieve all grid cells within a square region with edge size 84 to 144 km around it, resulting in approximately 50 to 150 entries in the 12-by-12 km grid. In the **tabular** modality, we prepare each table with numerical values, a caption, and row and column indices as textual strings. In the **image** modality, we prepare three types of visualization with increasing information densities, as shown in Figure 2: (1) A standalone heatmap, (2) A heatmap with overlaid numerical annotations at each grid cell. and (3) A heatmap overlaid on an actual geographic base map. Specifically, we render the tabular data as a heatmap with color gradients. This heatmap is optionally added with numerical annotation of the value on each cell, or overlaid on a base map (OpenStreetMap contributors, 2024) using Folium (Folium, 2023). To maintain consistency with the tabular format, we also render row and column indices around the heatmap. This visualization offers a richer representation to mirror common practices in real-world analysis. To isolate the challenge of data retrieval, GeoGrid-Bench provides the foundation model during evaluation with all necessary data frames in either tabular or image formats, focusing solely on whether the model can solve the problem given the relevant information.

GeoGrid-Bench builds on expert-curated templates for scalable query generation To ensure that the benchmark reflects the types of analysis most relevant to practitioners in geospatial research, we consulted 13 domain experts. These experts routinely engage with geo-spatial data to identify patterns, assess risks, and support decision-making under uncertainty. We develop eight representative question templates based on operational needs identified by experts. Each template takes as input one or two climate variables, locations, and time frames and outputs a filled-in user query in our benchmark, and may require between one and eight data frames to answer. This structured approach enables the automatic generation of a wide variety of concrete, data-driven queries. For every template, we manually craft oracle code that deterministically solves the question and prepares ground-truth answers in desired formats. *Crucially, the same oracle applies uniformly to every query generated from a given template, enabling the scalability of the benchmark. As a result, once a template and its oracle are validated, we ensure the quality of every generated instance.*

Each question is a multiple-choice with four options, all generated by the oracle code rather than a language model. Recognizing that a foundation model may excel at different aspects in answering a geo-spatial query, the benchmark has each query probe a different aspect in giving the answer, as shown in Figure 1. Specifically, answer options target the following aspects: (1) Overall patterns (e.g., the wildfire risk overall increases). (2) Spatial references (e.g., the highest wildfire risk occurs around the top-left region). (3) Coordinate references (e.g., the highest wildfire risk occurs around Column 204 Row 106). (4) Label references (e.g., the highest wildfire risk occurs near the textual label "Santa Clara" on the map), which is only available for the image type "heatmap overlaid on an actual geographic base map".

In addition, to explore which data modalities most effectively support geo-spatial analysis, we evaluate models across three input settings: **language-only**, **language and code**, and **language and vision**. Detailed prompting strategies for each setting are provided in Appendix A. In each mode, we provide the model with the user query, the relevant data (in either tabular or image format), all four multiple-choice options, and system instructions as inputs. We extract the model's final answer following the special tokens *"####Final Answer"* to facilitate answer parsing. If the model fails to provide an explicit option (a), (b), (c), or (d), we use a sentence embedding model (Reimers & Gurevych, 2019) to identify the most similar option based on the model's response. When the model outputs Python code, we execute the code in a shell environment to extract the final answers.

4 Experiment

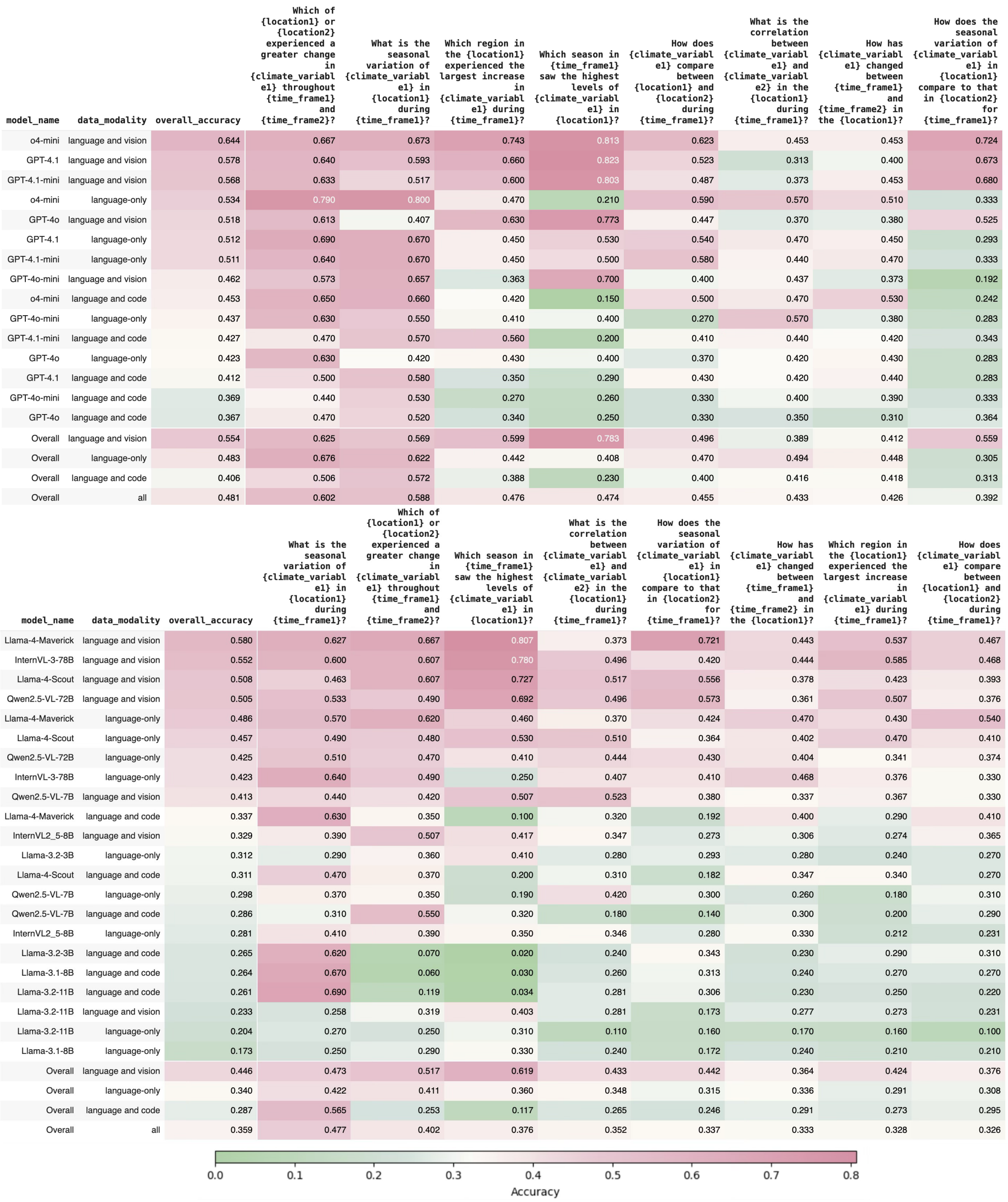
4.1 Experimental Setup

We benchmark a range of state-of-the-art closed-source and open-source models on GeoGrid-Bench. Our evaluation covers 5 models from OpenAI, including o4-mini, GPT-4.1, GPT-4.1-mini, GPT-4o, and GPT-4o-mini (OpenAI, 2024, 2025; Hurst et al., 2024), and 6 open-source models including Llama-4-Maverick, Llama-4-Scout, Llama-3.2-11B-Vision, Llama-3.2-3B, Llama-3.1-8B (Grattafiori et al., 2024; AI, 2024), and Qwen-2.5-VL-7B (Bai et al., 2025). OpenAI models are accessed via API calls, and Llama-4 models are accessed through the Lambda Inference API. Inferences for other open-source models run locally on four NVIDIA A100-SXM4 GPUs with 40GB of VRAM. For all models, we set `max_new_tokens` as 1024 with default temperature and sampling strategies. To ensure fair evaluation across all models, we used identical zero-shot prompts for every model tested on a randomly sampled subset of 3,200 examples from GeoGrid-Bench. We conducted additional ablation experiments using 3-shot prompts on 2 representative open-source models (see Appendix C).

4.2 Evaluation Results and Findings

Vision-language models achieve the strongest performance in geo-spatial tasks Among the models we evaluate, o4-mini achieves overall the highest performance, while Llama-4-Maverick leads among open-source models, as shown in Figure 4. Overall, models that receive input in the vision modality consistently outperform those using language-only input. This suggests that converting geo-spatial gridded data into heatmap visualizations—rather than presenting models directly with large volumes of raw numerical values in tabular forms—enables foundation models to more effectively interpret such data with complex spatial-temporal patterns. Statistical analyses confirmed no systematic geographic or temporal biases across the evaluated models (see Appendix D).

Inferior performance in code highlights the need for more agentic models in geo-spatial tasks Contrary to our expectations, foundation models leveraging programming code do not outperform their language-only counterparts on our task. Upon closer inspection, much of the generated code is not directly executable in a single pass. For instance, models produce incomplete scripts or bugs, omit expected outputs, fail to parse data, or struggle with planning over geo-spatial data—ultimately requiring human intervention across multiple iterations. This limitation aligns with how we construct the oracle code in the benchmark. This issue is more severe in open-source models like Llama, which tend to produce fewer executable code. We, therefore, emphasize the need for stronger *agentic* behaviors (Plaat et al., 2025; Kapoor et al., 2024; Ng, 2024) in foundation models, where we define "agentic" as the ability to autonomously generate fully executable code for human end-users in a single interaction, particularly when the end-users are domain scientists rather than programmers.



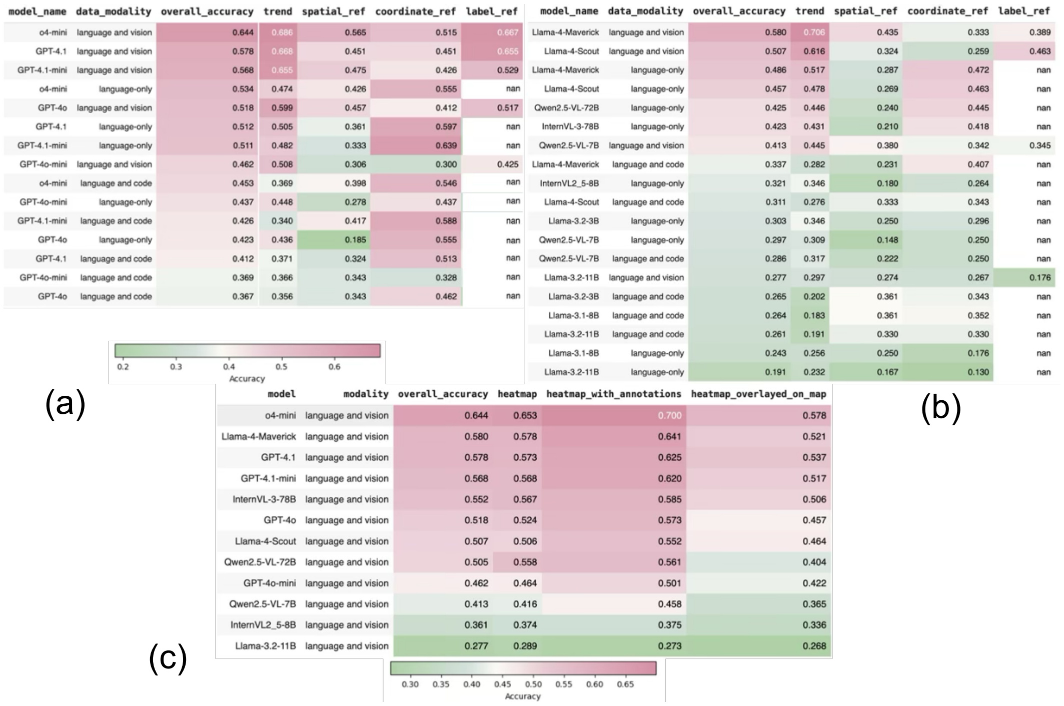
model_name	data_modality	overall_accuracy	Which of {location1} or {location2} experienced a greater change in {climate_variable1} throughout {time_frame1} and {time_frame2}?	What is the seasonal variation of {climate_variable1} in {location1} during {time_frame1}?	Which region in the {location1} experienced the largest increase in {climate_variable1} during {time_frame1}?	Which season in {time_frame1} saw the highest levels of {climate_variable1} in {location1}?	How does {climate_variable1} compare between {location1} and {location2} during {time_frame1}?	What is the correlation between {climate_variable1} and {climate_variable2} in the {location1} during {time_frame1}?	How has {climate_variable1} changed between {time_frame1} and {time_frame2} in the {location1}?	How does the seasonal variation of {climate_variable1} in {location1} compare to that in {location2} for {time_frame1}?
o4-mini	language and vision	0.644	0.667	0.673	0.743	0.813	0.623	0.453	0.453	0.724
GPT-4.1	language and vision	0.578	0.640	0.593	0.660	0.823	0.523	0.313	0.400	0.673
GPT-4.1-mini	language and vision	0.568	0.633	0.517	0.600	0.803	0.487	0.373	0.453	0.680
o4-mini	language-only	0.534	0.790	0.800	0.470	0.210	0.590	0.570	0.510	0.333
GPT-4o	language and vision	0.518	0.613	0.407	0.630	0.773	0.447	0.370	0.380	0.525
GPT-4.1	language-only	0.512	0.690	0.670	0.450	0.530	0.540	0.470	0.450	0.293
GPT-4.1-mini	language-only	0.511	0.640	0.670	0.450	0.500	0.580	0.440	0.470	0.333
GPT-4o-mini	language and vision	0.462	0.573	0.657	0.363	0.700	0.400	0.437	0.373	0.192
o4-mini	language and code	0.453	0.650	0.660	0.420	0.150	0.500	0.470	0.530	0.242
GPT-4o-mini	language-only	0.437	0.630	0.550	0.410	0.400	0.270	0.570	0.380	0.283
GPT-4.1-mini	language and code	0.427	0.470	0.570	0.560	0.200	0.410	0.440	0.420	0.343
GPT-4o	language-only	0.423	0.630	0.420	0.430	0.400	0.370	0.420	0.430	0.283
GPT-4.1	language and code	0.412	0.500	0.580	0.350	0.290	0.430	0.420	0.440	0.283
GPT-4o-mini	language and code	0.369	0.440	0.530	0.270	0.260	0.330	0.400	0.390	0.333
GPT-4o	language and code	0.367	0.470	0.520	0.340	0.250	0.330	0.350	0.310	0.364
Overall	language and vision	0.554	0.625	0.569	0.599	0.783	0.496	0.389	0.412	0.559
Overall	language-only	0.483	0.676	0.622	0.442	0.408	0.470	0.494	0.448	0.305
Overall	language and code	0.406	0.506	0.572	0.388	0.230	0.400	0.416	0.418	0.313
Overall	all	0.481	0.602	0.588	0.476	0.474	0.455	0.433	0.426	0.392

model_name	data_modality	overall_accuracy	What is the seasonal variation of {climate_variable1} in {location1} during {time_frame1}?	Which of {location1} or {location2} experienced a greater change in {climate_variable1} throughout {time_frame1} and {time_frame2}?	Which season in {time_frame1} saw the highest levels of {climate_variable1} in {location1}?	What is the correlation between {climate_variable1} and {climate_variable2} in the {location1} during {time_frame1}?	How does the seasonal variation of {climate_variable1} in {location1} compare to that in {location2} for {time_frame1}?	How has {climate_variable1} changed between {time_frame1} and {time_frame2} in the {location1}?	Which region in the {location1} experienced the largest increase in {climate_variable1} during {time_frame1}?	How does {climate_variable1} compare between {location1} and {location2} during {time_frame1}?
Llama-4-Maverick	language and vision	0.580	0.627	0.667	0.807	0.373	0.721	0.443	0.537	0.467
InternVL-3-78B	language and vision	0.552	0.600	0.607	0.780	0.496	0.420	0.444	0.585	0.468
Llama-4-Scout	language and vision	0.508	0.463	0.607	0.727	0.517	0.556	0.378	0.423	0.393
Qwen2.5-VL-72B	language and vision	0.505	0.533	0.490	0.692	0.496	0.573	0.361	0.507	0.376
Llama-4-Maverick	language-only	0.486	0.570	0.620	0.460	0.370	0.424	0.470	0.430	0.540
Llama-4-Scout	language-only	0.457	0.490	0.480	0.530	0.510	0.364	0.402	0.470	0.410
Qwen2.5-VL-72B	language-only	0.425	0.510	0.470	0.410	0.444	0.430	0.404	0.341	0.374
InternVL-3-78B	language-only	0.423	0.640	0.490	0.250	0.407	0.410	0.468	0.376	0.330
Qwen2.5-VL-7B	language and vision	0.413	0.440	0.420	0.507	0.523	0.380	0.337	0.367	0.330
Llama-4-Maverick	language and code	0.337	0.630	0.350	0.100	0.320	0.192	0.400	0.290	0.410
InternVL2_5-8B	language and vision	0.329	0.390	0.507	0.417	0.347	0.273	0.306	0.274	0.365
Llama-3.2-3B	language-only	0.312	0.290	0.360	0.410	0.280	0.293	0.280	0.240	0.270
Llama-4-Scout	language and code	0.311	0.470	0.370	0.200	0.310	0.182	0.347	0.340	0.270
Qwen2.5-VL-7B	language-only	0.298	0.370	0.350	0.190	0.420	0.300	0.260	0.180	0.310
Qwen2.5-VL-7B	language and code	0.286	0.310	0.550	0.320	0.180	0.140	0.300	0.200	0.290
InternVL2_5-8B	language-only	0.281	0.410	0.390	0.350	0.346	0.280	0.330	0.212	0.231
Llama-3.2-3B	language and code	0.265	0.620	0.070	0.020	0.240	0.343	0.230	0.290	0.310
Llama-3.1-8B	language and code	0.264	0.670	0.060	0.030	0.260	0.313	0.240	0.270	0.270
Llama-3.2-11B	language and code	0.261	0.690	0.119	0.034	0.281	0.306	0.230	0.250	0.220
Llama-3.2-11B	language and vision	0.233	0.258	0.319	0.403	0.281	0.173	0.277	0.273	0.231
Llama-3.2-11B	language-only	0.204	0.270	0.250	0.310	0.110	0.160	0.170	0.160	0.100
Llama-3.1-8B	language-only	0.173	0.250	0.290	0.330	0.240	0.172	0.240	0.210	0.210
Overall	language and vision	0.446	0.473	0.517	0.619	0.433	0.442	0.364	0.424	0.376
Overall	language-only	0.340	0.422	0.411	0.360	0.348	0.315	0.336	0.291	0.308
Overall	language and code	0.287	0.565	0.253	0.117	0.265	0.246	0.291	0.273	0.295
Overall	all	0.359	0.477	0.402	0.376	0.352	0.337	0.333	0.328	0.326

Figure 4: Evaluation results. The top table shows OpenAI models and the bottom table shows open-source models. Each row corresponds to one model with one data modality—language-only, language and code, or language and vision, while each column represents a query template in Table 1.

Common error patterns in geo-spatial reasoning We randomly collected 50 examples where models produced incorrect answers and identified several common error patterns. First, models sometimes provided step-by-step analytical plans without converting them into explicit mathematical calculations, instead giving final answers directly after the plan. Second, some analyses failed to extract actual values from the provided data tables and instead relied on the model's own assumptions rather than the actual data. Third, models sometimes focused on and were distracted by local regional patterns rather than analyzing overall correlations across the spatial domain. Finally, when visualizations were provided, models occasionally failed to extract relevant textual annotations and numerical markers from the images, limiting their ability to perform precise quantitative analysis.

Fine-grained geo-spatial tasks reveals different strength-weakness tradeoffs Commercial and open-source models exhibit different strengths and weaknesses in fine-grained geo-spatial tasks, as



(a)

model_name	data_modality	overall_accuracy	trend	spatial_ref	coordinate_ref	label_ref
o4-mini	language and vision	0.644	0.686	0.565	0.515	0.667
GPT-4.1	language and vision	0.578	0.668	0.451	0.451	0.655
GPT-4.1-mini	language and vision	0.568	0.655	0.475	0.426	0.529
o4-mini	language-only	0.534	0.474	0.426	0.555	nan
GPT-4o	language and vision	0.518	0.599	0.457	0.412	0.517
GPT-4.1	language-only	0.512	0.505	0.361	0.597	nan
GPT-4.1-mini	language-only	0.511	0.482	0.333	0.639	nan
GPT-4o-mini	language and vision	0.462	0.508	0.306	0.300	0.425
o4-mini	language and code	0.453	0.369	0.398	0.546	nan
GPT-4o-mini	language-only	0.437	0.448	0.278	0.437	nan
GPT-4.1-mini	language and code	0.426	0.340	0.417	0.588	nan
GPT-4o	language-only	0.423	0.436	0.185	0.555	nan
GPT-4.1	language and code	0.412	0.371	0.324	0.513	nan
GPT-4o-mini	language and code	0.369	0.366	0.343	0.328	nan
GPT-4o	language and code	0.367	0.356	0.343	0.462	nan

(b)

model_name	data_modality	overall_accuracy	trend	spatial_ref	coordinate_ref	label_ref
Llama-4-Maverick	language and vision	0.580	0.706	0.435	0.333	0.389
Llama-4-Scout	language and vision	0.507	0.616	0.324	0.259	0.463
Llama-4-Maverick	language-only	0.486	0.517	0.287	0.472	nan
Llama-4-Scout	language-only	0.457	0.478	0.269	0.463	nan
Qwen2.5-VL-72B	language-only	0.425	0.446	0.240	0.445	nan
InternVL-3-78B	language-only	0.423	0.431	0.210	0.418	nan
Qwen2.5-VL-7B	language and vision	0.413	0.445	0.380	0.342	0.345
Llama-4-Maverick	language and code	0.337	0.282	0.231	0.407	nan
InternVL2_5-8B	language-only	0.321	0.346	0.180	0.264	nan
Llama-4-Scout	language and code	0.311	0.276	0.333	0.343	nan
Llama-3.2-3B	language-only	0.303	0.346	0.250	0.296	nan
Qwen2.5-VL-7B	language-only	0.297	0.309	0.148	0.250	nan
Qwen2.5-VL-7B	language and code	0.286	0.317	0.222	0.250	nan
Llama-3.2-11B	language and vision	0.277	0.297	0.274	0.267	0.176
Llama-3.2-3B	language and code	0.265	0.202	0.361	0.343	nan
Llama-3.1-8B	language and code	0.264	0.183	0.361	0.352	nan
Llama-3.2-11B	language and code	0.261	0.191	0.330	0.330	nan
Llama-3.1-8B	language-only	0.243	0.256	0.250	0.176	nan
Llama-3.2-11B	language-only	0.191	0.232	0.167	0.130	nan

(c)

model	modality	overall_accuracy	heatmap	heatmap_with_annotations	heatmap_overlayed_on_map
o4-mini	language and vision	0.644	0.653	0.700	0.578
Llama-4-Maverick	language and vision	0.580	0.578	0.641	0.521
GPT-4.1	language and vision	0.578	0.573	0.625	0.537
GPT-4.1-mini	language and vision	0.568	0.568	0.620	0.517
InternVL-3-78B	language and vision	0.552	0.567	0.585	0.506
GPT-4o	language and vision	0.518	0.524	0.573	0.457
Llama-4-Scout	language and vision	0.507	0.506	0.552	0.464
Qwen2.5-VL-72B	language and vision	0.505	0.558	0.561	0.404
GPT-4o-mini	language and vision	0.462	0.464	0.501	0.422
Qwen2.5-VL-7B	language and vision	0.413	0.416	0.458	0.365
InternVL2_5-8B	language and vision	0.361	0.374	0.375	0.336
Llama-3.2-11B	language and vision	0.277	0.289	0.273	0.268

Figure 5: More evaluation results. (a) OpenAI models and (b) open-source models evaluated under different data modalities. Columns represent fine-grained answer aspects defined in Section 3, including trend, spatial references, coordinate references, and label references. There exist NaN values since the label reference is only available for the vision modality. (c) vision-language models, which are evaluated on three visualization types, as mentioned in Section 3 and Figure 2.

shown in Figure 4. Specifically, open-source models generally struggle more than commercial ones in identifying regions with the most significant patterns. However, both types of models perform well when comparing trends between two locations or analyzing seasonal variations at a single location. In contrast, they show weaker performance when comparing seasonal variations across multiple locations or comparing data across different locations and times.

Models perform better at identifying overall trends than fine-grained region detections As mentioned in Figure 1, target answers captures fine-grained aspects in answering these geo-spatial queries. Evaluation results in Figure 5 (a) and (b) show that models perform best on the "trend" column, while accuracy drops for spatial, coordinate, or label references—highlighting a need for improvement in fine-grained regional understanding.

Heatmaps with numerical annotations enhance performance, whereas map-overlaid heatmaps pose greater challenges for vision-language models Figure 5 (c) compares model performance across three input image formats defined in Figure 2. Adding numerical annotations to heatmaps improves model accuracy compared to using color gradients alone. In contrast, the most realistic format, where heatmaps are overlaid on geographic base maps, poses the greatest challenge for all models, as the added visual complexity hinders spatial pattern recognition.

5 Related Work

Geo-Spatial Reasoning with LLMs Geo-spatial reasoning involves understanding and analyzing complex data based on its spatial and temporal relationships in the world (Schottlander & Shekel, 2025). Most existing work focuses on Earth observation data from satellite or remote sensing imagery (Lacoste et al., 2023; Zhang et al., 2023b; Danish et al., 2024; Zhang & Wang, 2024; Zheng et al., 2023; Wang et al., 2024a; Muhtar et al., 2024; Bazi et al., 2024; Kuckreja et al., 2024; Tao et al., 2025; Liu et al., 2025), performing scene understanding tasks such as object detection, semantic segmentation, object counting, captioning. Notable examples include GeoGPT (Zhang et al., 2023b), GeoBench (Danish et al., 2024), EarthVQA (Wang et al., 2024a), GEOBench-VLM (Danish et al., 2024), and GeoChat (Kuckreja et al., 2024). However, gridded geo-spatial data is critical for capturing spatial and temporal patterns, remains largely overlooked in the AI-assisted geo-spatial research. Our work, GeoGrid-Bench, specifically targets this gap by focusing on grid-based data in both tabular and image formats, and evaluating how foundation models can analyze the underlying patterns. Other efforts in geo-spatial research have focused on text-based data retrieval with tool usages, particularly through Geographic Information Systems (GIS) (National Geographic Society, 2025), SQL, or GeoSPARQL (van Rees, 2013) queries (Krechetova & Kochedykov, 2025; Ning et al., 2025; Mooney et al., 2023; Li & Ning, 2023; Jiang & Yang, 2024; Resch et al., 2025; Zhang et al., 2023b; Jiang et al., 2024e) or Retrieval-Augmented Generation (RAG) systems (Cromp et al., 2024; Xie et al., 2024, 2025; Vaghefi et al., 2023; Thulke et al., 2024; Bulian et al., 2023). Representative works include GeoGPT (Zhang et al., 2023b), GeoBenchX (Krechetova & Kochedykov, 2025), Autonomous GIS (Li & Ning, 2023), WildfireGPT (Xie et al., 2024, 2025), and ChatClimate Vaghefi et al. (2023). These approaches typically present geo-spatial information in textual formats and then rely on specific query syntax or semantic embeddings to interact with their databases. In contrast, our work sidesteps the data retrieval part and focuses on the geo-spatial data analysis itself.

Tabular Reasoning with LLMs Gridded geo-spatial data is often represented in tabular formats, posing unique challenges for language models in processing structured, numerically dense information. Current literature primarily focus on tables from databases with rich semantic annotations such as a descriptive name of each entity. Benchmarks like HybridQA, TabFact, ToTTo, WikiTQ, and others (Chen et al., 2020, 2019; Parikh et al., 2020; Aly et al., 2021; Chen et al., 2021; Pasupat & Liang, 2015) focus on simple fact extractions and He et al. (2024b); Sui et al. (2024) cover more advanced analysis that still rely heavily on semantic cues. In contrast, our work focuses on tables dominated by large volumes of numerical values, with spatial dependencies and no semantic annotations except for coordinates, presenting a different form of tabular reasoning (Fang et al., 2024; Zhang et al., 2025). To handle tabular data with language models, current work adopts strategies such as serializing tables into Markdown or other common formats (Fang et al., 2024; Wang et al., 2024b), fine-tuning on tabular tasks (Yang et al., 2023; Zhang et al., 2023a; Li et al., 2023; Thomas et al., 2024), leveraging tool use and code generation (Fang et al., 2024; Cromp et al., 2024; Cheng et al., 2022; Zhang et al., 2023c), or using image-based table representations (Deng et al., 2024). In our work, we extend this line of research by visualizing tables with geo-spatial semantics heatmaps or overlays on actual maps and by exploring code-based analysis in geo-spatial contexts that introduce unique challenges.

6 Conclusion

We introduced GeoGrid-Bench, a comprehensive benchmark designed to evaluate the capability of foundation models to understand multimodal gridded geo-spatial data. GeoGrid-Bench features structured, dense numerical data using real-world gridded datasets and expert-curated templates to evaluate scientifically relevant geo-spatial tasks. This integrated design enables robust and scalable assessment of foundation models across vision, language, and code modalities. Our evaluation reveals that while vision-language models excel at interpreting spatial patterns from heatmaps, they still struggle with fine-grained regional understanding and label-based reasoning. Meanwhile, language and code models show limited success in generating executable analysis scripts without human intervention, highlighting the need for stronger agentic behavior. These findings point to several critical areas where model capabilities must improve to meet the practical needs of geo-spatial scientific analysis. Overall, this work can inform the development of more capable models to process and understand the dense numerical data, spatiotemporal dependencies, and multimodal representations of geo-spatial data, supporting the advancement of foundation models for informed decision-making and resilience building across a wide range of real-world challenges.

Limitations and Future Work We acknowledge that this dataset is limited to the United States due to data availability. Additionally, our benchmark focuses on geo-spatial data in gridded formats, intentionally excluding other common data types such as Earth observation and remote sensing imagery, which have already been extensively studied in prior work. However, the underlying framework are designed to be generalizable and can be readily applied to similar gridded geo-spatial datasets from other regions. Building on this foundation, future work will focus on expanding GeoGrid-Bench beyond the United States and incorporating richer data modalities such as satellite imagery, elevation maps, and land use data to enable broader and more diverse analytical capabilities.

References

Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation, 2024. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed: 2025-04-27.

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. Feverous: Fact extraction and verification over unstructured and structured information. *arXiv preprint arXiv:2106.05707*, 2021.

Argonne National Laboratory. Climrr: Climate risk and resilience portal. https://climrr.anl.gov, 2023. Accessed: 2025-04-15.

Argonne National Laboratory. Geospatial Energy Mapper, 2025. URL https://gem.anl.gov/.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Vidhisha Balachandran, Jingya Chen, Neel Joshi, Besmira Nushi, Hamid Palangi, Eduardo Salinas, Vibhav Vineet, James Woffinden-Luey, and Safoora Yousefi. Eureka: Evaluating and understanding large foundation models. *arXiv preprint arXiv:2409.10566*, 2024.

Yakoub Bazi, Laila Bashmal, Mohamad Mahmoud Al Rahhal, Riccardo Ricci, and Farid Melgani. Rs-llava: A large vision-language model for joint captioning and question answering in remote sensing imagery. *Remote Sensing*, 16(9):1477, 2024.

Prabin Bhandari, Antonios Anastasopoulos, and Dieter Pfoser. Are large language models geospatially knowledgeable? In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pp. 1–4, 2023.

Jannis Bulian, Mike S Schäfer, Afra Amini, Heidi Lam, Massimiliano Ciaramita, Ben Gaiarin, Michelle Chen Huebscher, Christian Buck, Niels G Mede, Markus Leippold, et al. Assessing large language models on climate information. *arXiv preprint arXiv:2310.02932*, 2023.

Center for Climate and Energy Solutions. Resilience Innovation Story: AT&T ClimRR, 2025. URL https://www.c2es.org/case-study/resilience-innovation-story-att-climrr/.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*, 2019.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Wang. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. *arXiv preprint arXiv:2004.07347*, 2020.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*, 2021.

Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, et al. Binding language models in symbolic languages. *arXiv preprint arXiv:2210.02875*, 2022.

Sonia Cromp, Behrad Rabiei, Maxwell Elling, Alexander Herron, and Michael Hendrickson. Climate pal: Climate analysis through conversational ai. 2024.

Muhammad Sohail Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. Geobench-vlm: Benchmarking vision-language models for geospatial tasks. *arXiv preprint arXiv:2411.19325*, 2024.

Maarten de Rijke, Bart van den Hurk, Flora Salim, Alaa Al Khourdajie, Nan Bai, Renato Calzone, Declan Curran, Getnet Demil, Lesley Frew, Noah Gießing, et al. Information retrieval for climate impact. *arXiv preprint arXiv:2504.01162*, 2025.

Naihao Deng, Zhenjie Sun, Ruiqi He, Aman Sikka, Yulong Chen, Lin Ma, Yue Zhang, and Rada Mihalcea. Tables as texts or images: Evaluating the table reasoning ability of llms and mllms. *arXiv preprint arXiv:2402.12424*, 2024.

Xi Fang, Weijie Xu, Fiona Anting Tan, Jiani Zhang, Ziqing Hu, Yanjun Qi, Scott Nickleach, Diego Socolinsky, Srinivasan Sengamedu, and Christos Faloutsos. Large language models (llms) on tabular data: Prediction, generation, and understanding–a survey. *arXiv preprint arXiv:2402.17944*, 2024.

Folium. Folium: Python data. leaflet.js maps. `https://python-visualization.github.io/folium/latest/`, 2023. URL `https://python-visualization.github.io/folium/latest/`. Version 0.14.0.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jiashu He, Mingyu Derek Ma, Jinxuan Fan, Dan Roth, Wei Wang, and Alejandro Ribeiro. Give: Structured reasoning with knowledge graph inspired veracity extrapolation. *arXiv preprint arXiv:2410.08475*, 2024a.

Xinyi He, Mengyu Zhou, Xinrun Xu, Xiaojun Ma, Rui Ding, Lun Du, Yan Gao, Ran Jia, Xu Chen, Shi Han, et al. Text2analysis: A benchmark of table question answering with advanced data analysis and unclear queries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18206–18215, 2024b.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Bowen Jiang, Yangxinyu Xie, Zhuoqun Hao, Xiaomeng Wang, Tanwi Mallick, Weijie J Su, Camillo J Taylor, and Dan Roth. A peek into token bias: Large language models are not yet genuine reasoners. *arXiv preprint arXiv:2406.11050*, 2024a.

Bowen Jiang, Yangxinyu Xie, Xiaomeng Wang, Yuan Yuan, Zhuoqun Hao, Xinyi Bai, Weijie J Su, Camillo J Taylor, and Tanwi Mallick. Towards rationality in language and multimodal agents: A survey. *arXiv preprint arXiv:2406.00252*, 2024b.

Bowen Jiang, Zhijun Zhuang, Shreyas S Shivakumar, Dan Roth, and Camillo J Taylor. Multi-agent vqa: Exploring multi-agent foundation models in zero-shot visual question answering. *arXiv preprint arXiv:2403.14783*, 2024c.

Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. A survey on large language models for code generation. *arXiv preprint arXiv:2406.00515*, 2024d.

Yongyao Jiang and Chaowei Yang. Is chatgpt a good geospatial data analyst? exploring the integration of natural language into structured query language within a spatial database. *ISPRS International Journal of Geo-Information*, 13(1):26, 2024.

Yue Jiang, Qin Chao, Yile Chen, Xiucheng Li, Shuai Liu, and Gao Cong. Urbanllm: Autonomous urban activity planning and management with large language models. *arXiv preprint arXiv:2406.12360*, 2024e.

Sayash Kapoor, Benedikt Stroebl, Zachary S Siegel, Nitya Nadgir, and Arvind Narayanan. Ai agents that matter. *arXiv preprint arXiv:2407.01502*, 2024.

Varvara Krechetova and Denis Kochedykov. Geobenchx: Benchmarking llms for multistep geospatial tasks. *arXiv preprint arXiv:2503.18129*, 2025.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27831–27840, 2024.

Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, et al. Geo-bench: Toward foundation models for earth monitoring. *Advances in Neural Information Processing Systems*, 36: 51080–51093, 2023.

Peng Li, Yeye He, Dror Yashar, Weiwei Cui, Song Ge, Haidong Zhang, Danielle Rifinski Fainman, Dongmei Zhang, and Surajit Chaudhuri. Table-gpt: Table-tuned gpt for diverse table tasks. *arXiv preprint arXiv:2310.09263*, 2023.

Zhenlong Li and Huan Ning. Autonomous gis: the next-generation ai-powered gis. *International Journal of Digital Earth*, 16(2):4668–4686, 2023.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023.

Zeping Liu, Fan Zhang, Junfeng Jiao, Ni Lao, and Gengchen Mai. Gair: Improving multimodal geo-foundation model with geo-aligned implicit representations. *arXiv preprint arXiv:2503.16683*, 2025.

Gengchen Mai, Chris Cundy, Kristy Choi, Yingjie Hu, Ni Lao, and Stefano Ermon. Towards a foundation model for geospatial artificial intelligence (vision paper). In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, pp. 1–4, 2022.

Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, et al. On the opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv preprint arXiv:2304.06798*, 2023.

Tanwi Mallick, Joshua David Bergerson, Duane R Verner, John K Hutchison, Leslie-Anne Levy, and Prasanna Balaprakash. Understanding the impact of climate change on critical infrastructure through nlp analysis of scientific literature. *Sustainable and Resilient Infrastructure*, 10(1):22–39, 2025.

Peter Mooney, Wencong Cui, Boyuan Guan, and Levente Juhász. Towards understanding the geospatial skills of chatgpt: Taking a geographic information systems (gis) exam. In *Proceedings of the 6th ACM SIGSPAtIAL international workshop on AI for geographic knowledge discovery*, pp. 85–94, 2023.

Dilxat Muhtar, Zhenshi Li, Feng Gu, Xueliang Zhang, and Pengfeng Xiao. Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. In *European Conference on Computer Vision*, pp. 440–457. Springer, 2024.

National Geographic Society. GIS (Geographic Information System). `https://education.nationalgeographic.org/resource/geographic-information-system-gis/`, 2025. Accessed: 2025-04-17.

Andrew Ng. Welcoming diverse approaches keeps machine learning strong. June 2024. URL `https://www.deeplearning.ai/the-batch/welcoming-diverse-approaches-keeps-machine-learning-strong/`.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

Huan Ning, Zhenlong Li, Temitope Akinboyewa, and M Naser Lessani. An autonomous gis agent framework for geospatial data retrieval. *International Journal of Digital Earth*, 18(1):2458688, 2025.

OpenAI. Openai o3 and o4-mini system card, 2024. URL `https://openai.com/index/o3-o4-mini-system-card/`. Accessed: 2025-04-18.

OpenAI. Introducing gpt-4.1 in the api, 2025. URL https://openai.com/index/gpt-4-1/. Accessed: 2025-05-12.

OpenStreetMap contributors. Openstreetmap, 2024. URL https://www.openstreetmap.org/. Accessed: 2025-05-12.

Ankur P Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. Totto: A controlled table-to-text generation dataset. *arXiv preprint arXiv:2004.14373*, 2020.

Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*, 2015.

Aske Plaat, Max van Duijn, Niki van Stein, Mike Preuss, Peter van der Putten, and Kees Joost Batenburg. Agentic large language models, a survey. *arXiv preprint arXiv:2503.23037*, 2025.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Bernd Resch, Polychronis Kolokoussis, David Hanny, Maria Antonia Brovelli, and Maged N Kamel Boulos. The generative revolution: Ai foundation models in geospatial health—applications, challenges and future research. *International Journal of Health Geographics*, 24:6, 2025.

David Schottlander and Tomer Shekel. Geospatial reasoning: Unlocking insights with generative ai and multiple foundation models. https://research.google/blog/geospatial-reasoning-unlocking-insights-with-generative-ai-and-multiple-foundation-models/, April 2025. Accessed: 2025-04-17.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pp. 645–654, 2024.

Lijie Tao, Haokui Zhang, Haizhao Jing, Yu Liu, Dawei Yan, Guoting Wei, and Xizhe Xue. Advancements in vision–language models for remote sensing: Datasets, capabilities, and enhancement techniques. *Remote Sensing*, 17(1):162, 2025.

TechBrew. ClimRR: Federal Climate Data Tool Helps Communities Plan for Extreme Weather, January 2025. URL https://www.techbrew.com/stories/2025/01/10/climrr-federal-climate-data-att.

Valentin Thomas, Junwei Ma, Rasa Hosseinzadeh, Keyvan Golestan, Guangwei Yu, Maks Volkovs, and Anthony L Caterini. Retrieval & fine-tuning for in-context tabular models. *Advances in Neural Information Processing Systems*, 37:108439–108467, 2024.

David Thulke, Yingbo Gao, Petrus Pelser, Rein Brune, Rricha Jalota, Floris Fok, Michael Ramos, Ian van Wyk, Abdallah Nasir, Hayden Goldstein, et al. Climategpt: Towards ai synthesizing interdisciplinary research on climate change. *arXiv preprint arXiv:2401.09646*, 2024.

Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Mathias Kraus, Simon Allen, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, et al. ChatClimate: Grounding conversational AI in climate science. *Communications Earth & Environment*, 4 (1):480, 2023.

Eric van Rees. Open geospatial consortium (ogc). *Geoinformatics*, 16(8):28, 2013.

Junjue Wang, Zhuo Zheng, Zihang Chen, Ailong Ma, and Yanfei Zhong. Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5481–5489, 2024a.

Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, et al. Chain-of-table: Evolving tables in the reasoning chain for table understanding. *arXiv preprint arXiv:2401.04398*, 2024b.

Yangxinyu Xie, Bowen Jiang, Tanwi Mallick, Joshua David Bergerson, John K Hutchison, Duane R Verner, Jordan Branham, M Ross Alexander, Robert B Ross, Yan Feng, et al. Wildfiregpt: Tailored large language model for wildfire analysis. *arXiv preprint arXiv:2402.07877*, 2024.

Yangxinyu Xie, Bowen Jiang, Tanwi Mallick, Joshua Bergerson, John K Hutchison, Duane R Verner, Jordan Branham, M Ross Alexander, Robert B Ross, Yan Feng, et al. Marsha: multi-agent rag system for hazard adaptation. *npj Climate Action*, 4(1):70, 2025.

Yazheng Yang, Yuqi Wang, Guang Liu, Ledell Wu, and Qi Liu. Unitabe: A universal pretraining protocol for tabular foundation model in data science. *arXiv preprint arXiv:2307.09249*, 2023.

Chenhui Zhang and Sherrie Wang. Good at captioning bad at counting: Benchmarking gpt-4v on earth observation data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7839–7849, 2024.

Tianshu Zhang, Xiang Yue, Yifei Li, and Huan Sun. Tablellama: Towards open large generalist models for tables. *arXiv preprint arXiv:2311.09206*, 2023a.

Xuanliang Zhang, Dingzirui Wang, Longxu Dou, Qingfu Zhu, and Wanxiang Che. A survey of table reasoning with large language models. *Frontiers of Computer Science*, 19(9):199348, 2025.

Yifan Zhang, Cheng Wei, Shangyou Wu, Zhengting He, and Wenhao Yu. Geogpt: Understanding and processing geospatial tasks through an autonomous gpt. *arXiv preprint arXiv:2307.07930*, 2023b.

Yunjia Zhang, Jordan Henkel, Avrilia Floratou, Joyce Cahoon, Shaleen Deep, and Jignesh M Patel. Reactable: enhancing react for table question answering. *arXiv preprint arXiv:2310.00815*, 2023c.

Haozhen Zheng, Chenhui Zhang, Kaiyu Guan, Yawen Deng, Sherrie Wang, Bruce L Rhoads, Andrew J Margenot, Shengnan Zhou, and Sheng Wang. Segment any stream: Scalable water extent detection with the segment anything model. In *NeurIPS 2023 Computational Sustainability: Promises and Pitfalls from Theory to Deployment*, 2023.

A Inference Prompts

To evaluate models across different modalities, we design prompts for three settings: language-only, language and code, and language and vision. Each prompt is designed to be simple yet encourage model response with desired style and consistent answer formatting.

- Language-only: models receive data in tabular format with instructions *"Think step by step before making a decision. Then, explicitly state your final choice after the special phrase "####Final Answer" followed by (a), (b), (c), or (d). Please don't use programming code."*.
- Language and programming code: models receive data in tabular format with instructions *"Please write Python code to answer the question and show the complete script. You must include a print statement at the end of the code that outputs the final answer using the special phrase "####Final Answer' followed by (a), (b), (c), or (d)."*
- Language and vision: models receive climate data in one of the three image formats with instructions *"Analyze this image and answer the question. Think step by step before making a decision. Then, explicitly state your final choice after the special phrase "####Final Answer" followed by (a), (b), (c), or (d)."*.

B Examples of Data Visualizations for All Query Templates

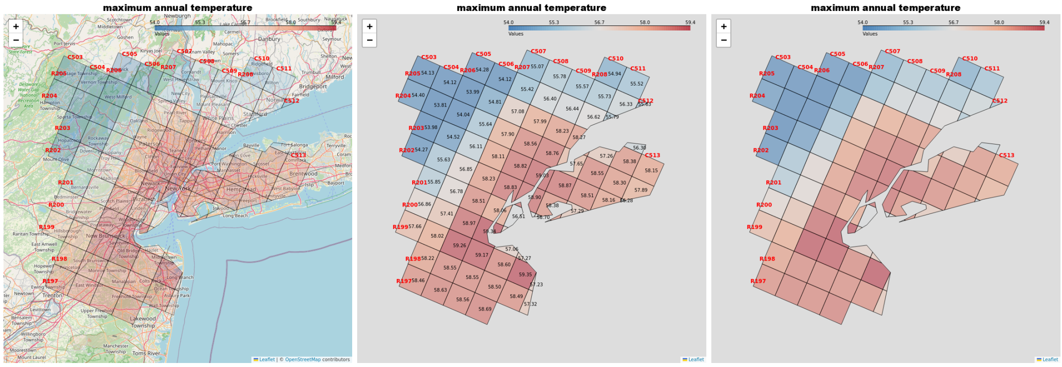


Figure 6: **Template 1:** Which region in {location1} experienced the largest increase in {climate_variable1} during {time_frame1}? This example takes location1 = New York City, NY, climate_variable1 = maximum annual temperate, and time_frame1 = historical period.

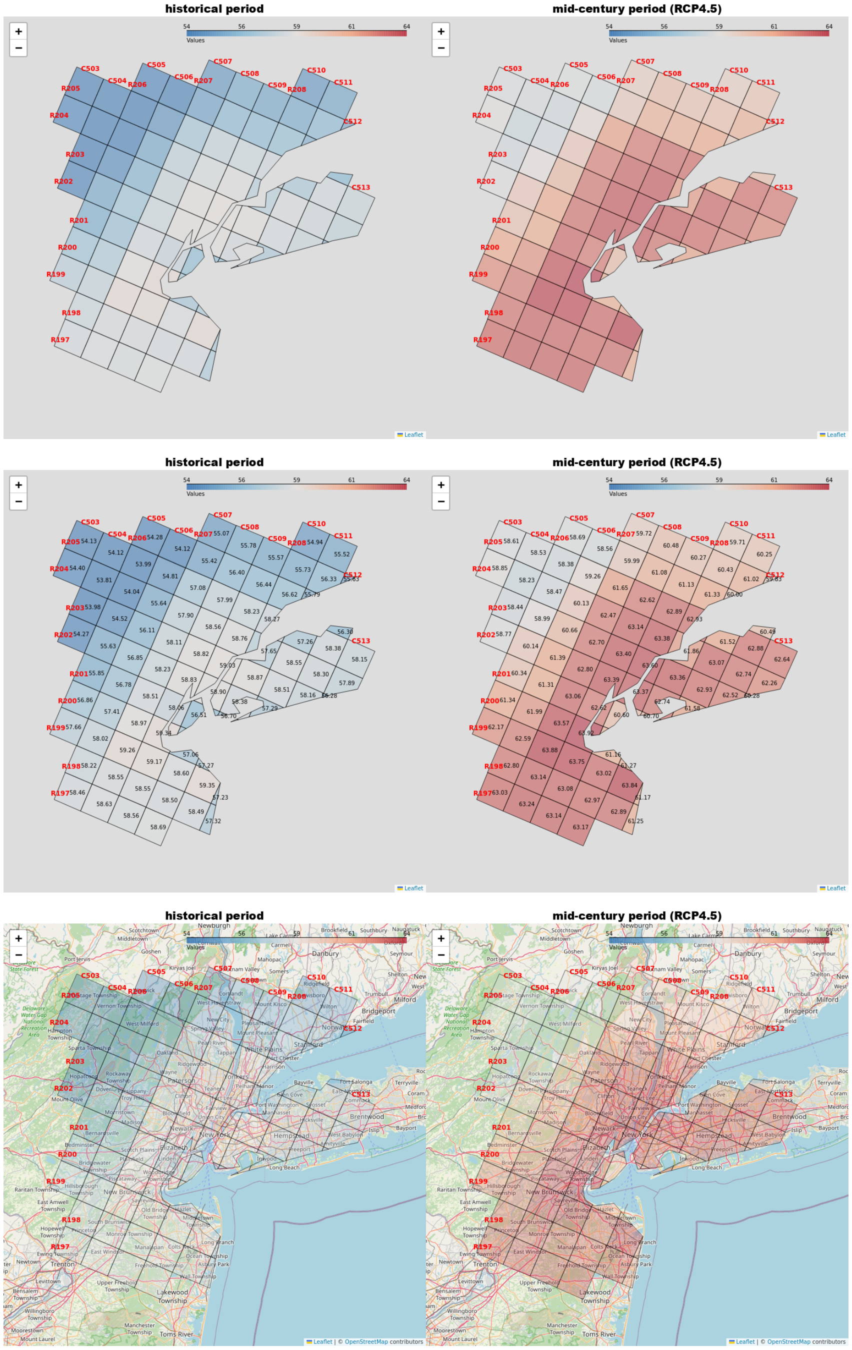


Figure 7: **Template 2:** How has {climate_variable1} changed between {time_frame1} and {time_frame2} in the {location1}? This example takes location1 = New York City, NY, climate_variable1 = maximum annual temperate, time_frame1 = historical period, and time_frame2 = mid-century period (RCP-4.5).

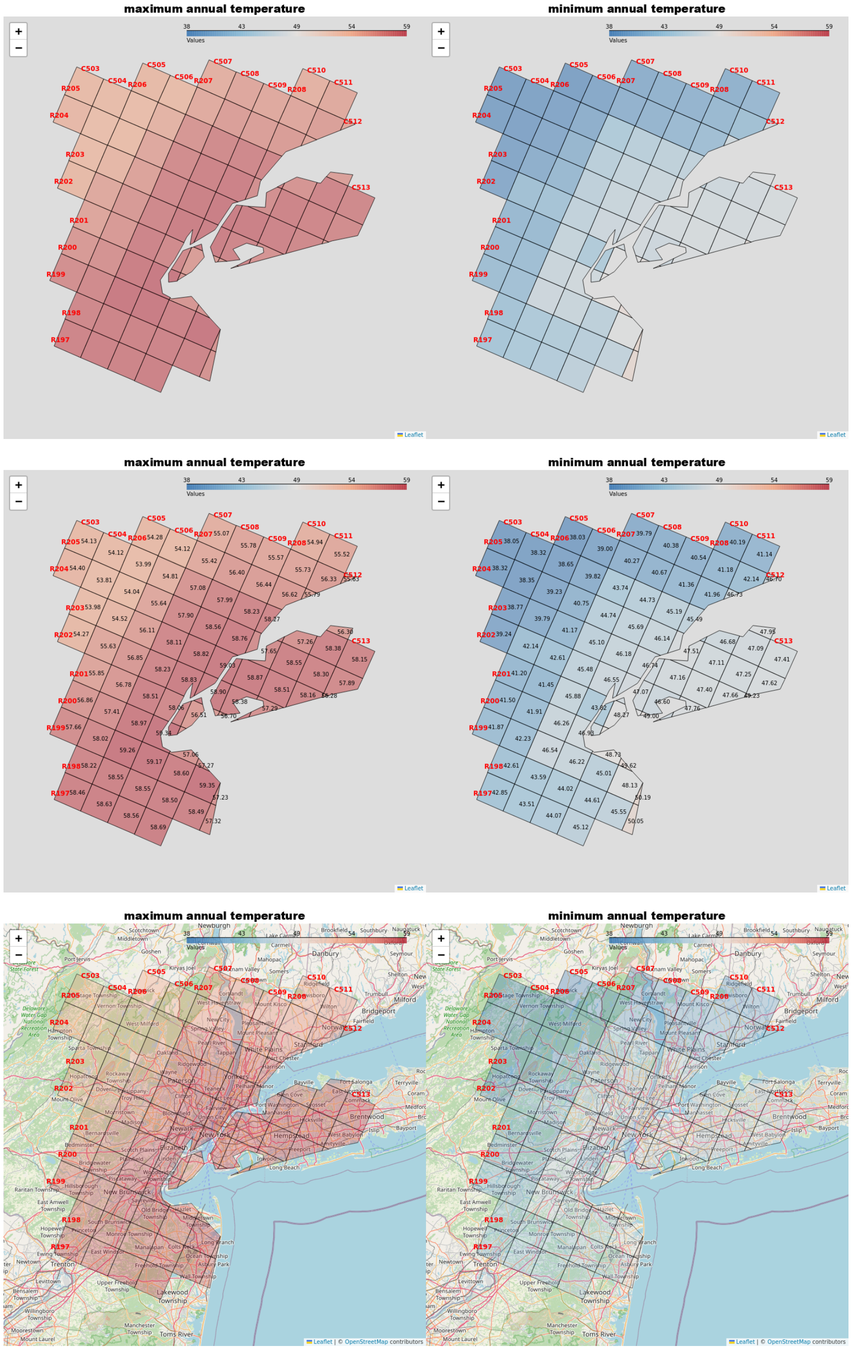


Figure 8: **Template 3:** What is the correlation between {climate_variable1} and {climate_variable2} in the {location1} during {time_frame1}? This example takes location1 = New York City, NY, climate_variable1 = maximum annual temperate, climate_variable2 = minimum annual temperate, and time_frame1 = historical period.

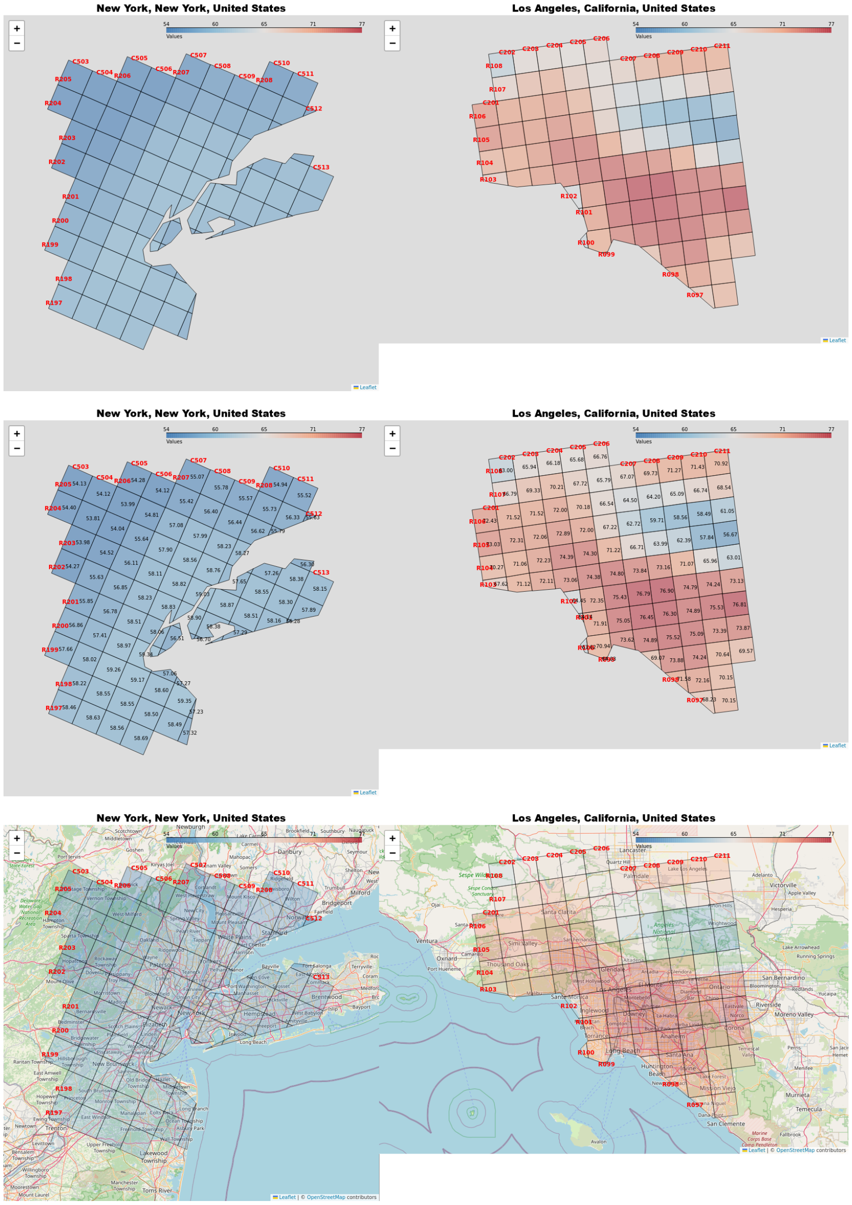


Figure 9: **Template 4:** How does {climate_variable1} compare between {location1} and {location2} during {time_frame1}? This example takes location1 = New York City, NY, location2 = Los Angeles, CA, climate_variable1 = maximum annual temperate, and time_frame1 = historical period.

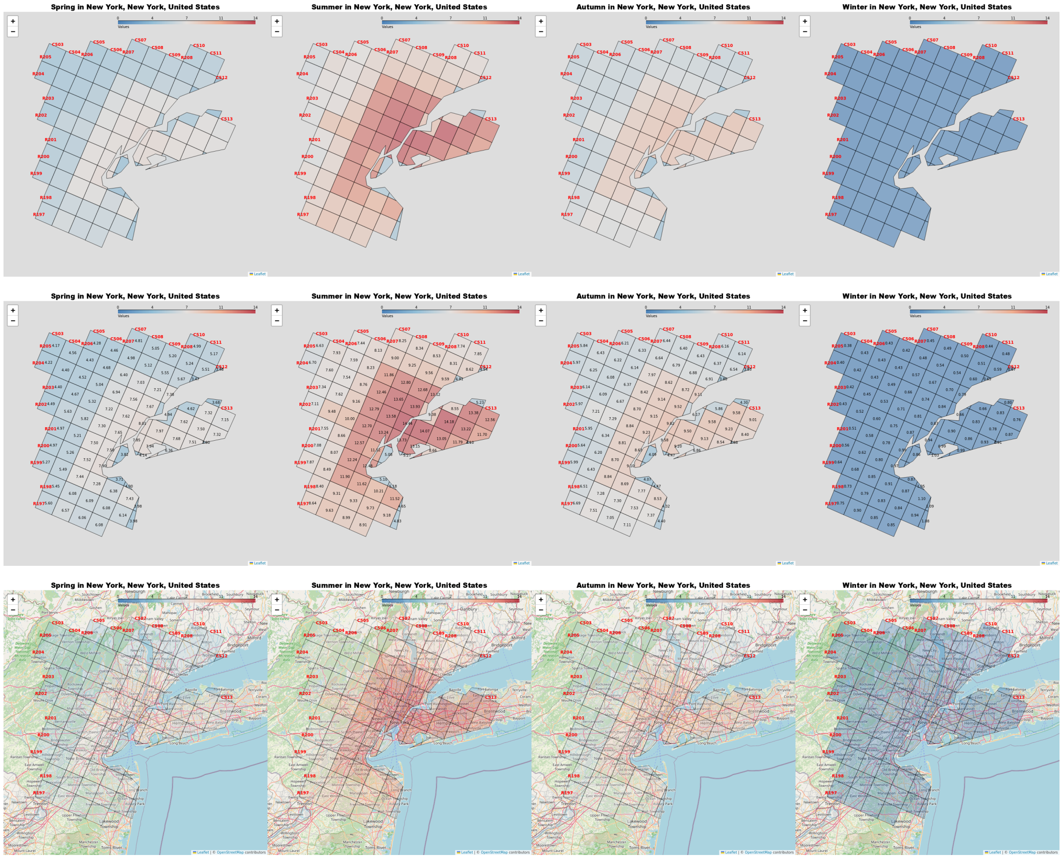
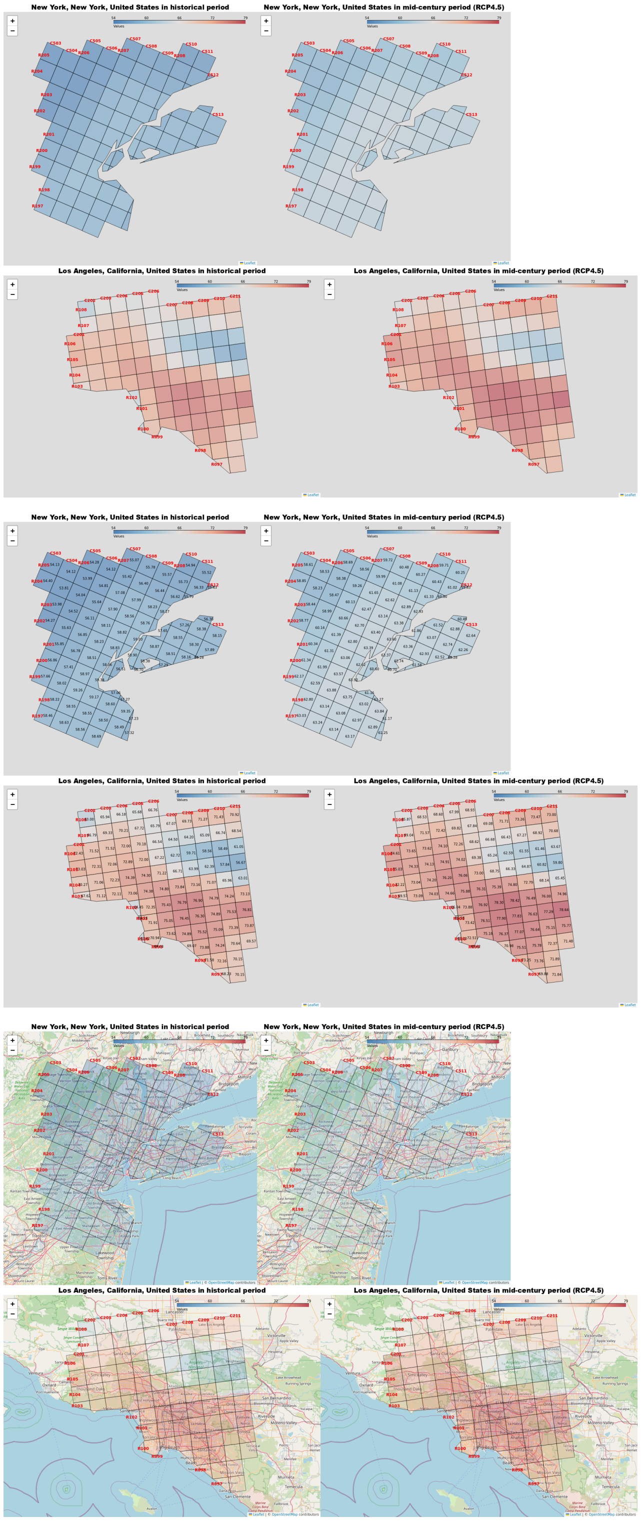

Spring in New York, New York, United States
Summer in New York, New York, United States
Autumn in New York, New York, United States
Winter in New York, New York, United States
Values
Leaflet

Figure 10: **Template 5**: What is the *seasonal* variation of {climate_variable1} in {location1} during {time_frame1}? Same data is used in **Template 6**: Which *season* in {time_frame1} saw the highest levels of {climate_variable1} in {location1}? This example takes location1 = New York City, NY, climate_variable1 = maximum annual temperate, and time_frame1 = historical period.


New York, New York, United States in historical period

New York, New York, United States in mid-century period (RCP4.5)

Los Angeles, California, United States in historical period

Los Angeles, California, United States in mid-century period (RCP4.5)

New York, New York, United States in historical period

New York, New York, United States in mid-century period (RCP4.5)

Los Angeles, California, United States in historical period

Los Angeles, California, United States in mid-century period (RCP4.5)

New York, New York, United States in historical period

New York, New York, United States in mid-century period (RCP4.5)

Los Angeles, California, United States in historical period

Los Angeles, California, United States in mid-century period (RCP4.5)

Figure 11: **Template 7.** Which of {location1} or {location2} experienced a greater change in {climate_variable1} throughout {time_frame1} and {time_frame2}? This example takes location1 = New York City, NY, location2 = Los Angeles, CA, climate_variable1 = maximum annual temperate, time_frame1 = historical period, and time_frame1 = mid-century period (RCP4.5).

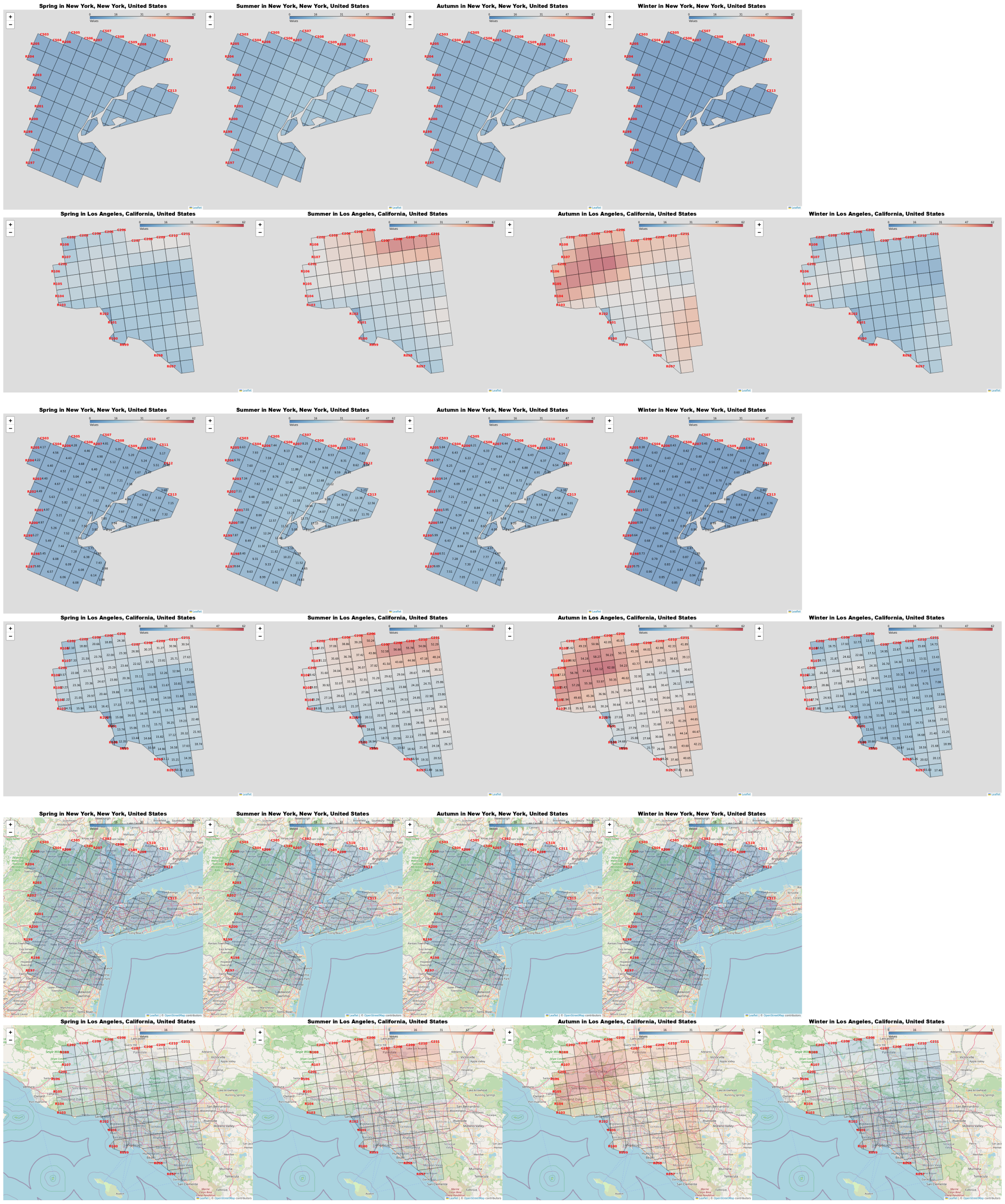

Spring in New York, New York, United States
Summer in New York, New York, United States
Autumn in New York, New York, United States
Winter in New York, New York, United States
Spring in Los Angeles, California, United States
Summer in Los Angeles, California, United States
Autumn in Los Angeles, California, United States
Winter in Los Angeles, California, United States
Spring in New York, New York, United States
Summer in New York, New York, United States
Autumn in New York, New York, United States
Winter in New York, New York, United States
Spring in Los Angeles, California, United States
Summer in Los Angeles, California, United States
Autumn in Los Angeles, California, United States
Winter in Los Angeles, California, United States
Spring in New York, New York, United States
Summer in New York, New York, United States
Autumn in New York, New York, United States
Winter in New York, New York, United States
Spring in Los Angeles, California, United States
Summer in Los Angeles, California, United States
Autumn in Los Angeles, California, United States
Winter in Los Angeles, California, United States

Figure 12: **Template 8.** How does the *seasonal* variation of {climate_variable1} in {location1} compare to that in {location2} for {time_frame1}? This example takes location1 = New York City, NY, location2 = Los Angeles, CA, climate_variable1 = maximum annual temperate, and time_frame1 = historical period.

C Ablation Study: Zero-shot vs. 3-shot Prompting

To assess the impact of prompting strategy on model performance, we conducted ablation experiments comparing zero-shot and 3-shot prompting approaches on two representative models: Llama-3.1-8B-Instruct and Qwen2.5-VL-7B-Instruct. Overall, results demonstrate performance improvements from zero-shot to 3-shot prompting.

Model Name	Prompting Strategy	Overall Accuracy
Llama-3.1-8B-Instruct	3-shot	0.196
Qwen2.5-VL-7B-Instruct	3-shot	0.369
Llama-3.1-8B-Instruct	zero-shot	0.173
Qwen2.5-VL-7B-Instruct	zero-shot	0.298

Table 2: Performance comparison between zero-shot and 3-shot prompting strategies on language-only tasks.

D Geographic and Temporal Bias Analysis

We conducted statistical analyses to test for geographic and temporal biases across four models: GPT-4o, GPT-4o-mini, Llama-4-Maverick-17b-128e, and Qwen2.5-VL-7b. We used one-way ANOVA tests with model accuracy as the dependent variable and geographic/temporal categories as independent variables. For geographic analysis, we grouped the questions by US regions (Northeast, South, Midwest, West) and city prominence (major vs. other cities). For temporal analysis, we categorized the questions by its relevance to historical, mid-century, and end-century periods. For GPT-4o (shown as example), the tests revealed no significant geographic bias across US regions (F=0.709, p=0.547), no temporal bias across historical/future periods (F=1.096, p=0.335), and no bias for more prominent cities (F=1.432, p=0.232). All four models failed to reject the null hypothesis across all tested dimensions. No systematic geographic or temporal bias exists in these models.