

Charting New Territories: Exploring the Geographic and Geospatial Capabilities of Multimodal LLMs

Jonathan Roberts
University of Cambridge
jdr53@cam.ac.uk

Timo Lüddecke
University of Göttingen
timo.lueddecke@uni-goettingen.de

Rehan Sheikh
University of Cambridge
rs2126@cam.ac.uk

Kai Han
The University of Hong Kong
kaihanx@hku.hk

Samuel Albanie
University of Cambridge
samuel.albanie.academic@gmail.com

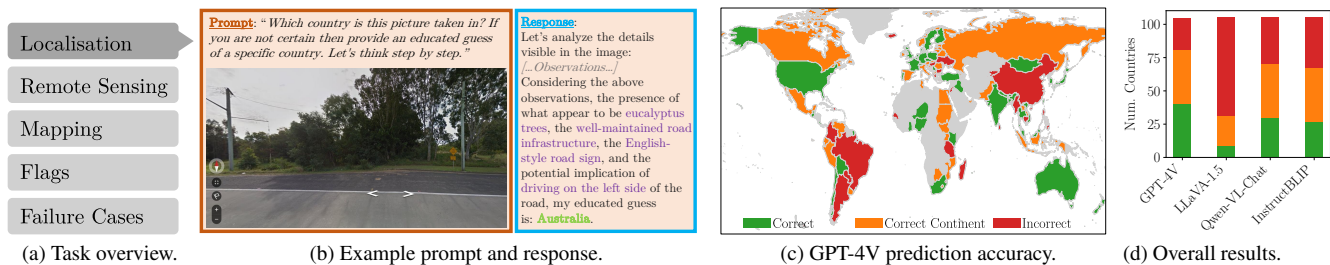


Figure 1. **Evaluating the geographic and geospatial capabilities of MLLMs.** (a) To probe MLLM knowledge we conduct a series of visual experiments. *Localisation*—(a) We pass each image to the MLLM together with the given prompt. As highlighted by its response, we find that GPT-4V is able to extract and reason over fine-grained details in such images. (c) GPT-4 attains promising accuracy on the benchmark across tested countries, where (d) it outperforms other strong MLLM baselines.

Abstract

Multimodal large language models (MLLMs) have shown remarkable capabilities across a broad range of tasks but their knowledge and abilities in the geographic and geospatial domains are yet to be explored, despite potential wide-ranging benefits to navigation, environmental research, and disaster response. We conduct a series of experiments exploring various vision capabilities of MLLMs within these domains, particularly focusing on the frontier model GPT-4V, and benchmark its performance against open-source counterparts. Our methodology involves challenging these models with a small-scale geographic benchmark consisting of a suite of visual tasks, testing their abilities across a spectrum of complexity. The analysis uncovers not only where such models excel, including instances where they outperform humans, but also where they falter, providing a balanced view of their capabilities in the geographic domain. To enable the comparison and evaluation of future models, we publicly release our benchmark.¹

¹Dataset available at <https://github.com/jonathan-roberts1/charting-new-territories>.

1. Introduction

Driven by key methodological advances (chiefly the Transformer architecture [31]) and larger-scale datasets and compute infrastructure, vision and language research are currently enjoying a golden age of advancement and progress. The growing capabilities of large language models (LLMs) [5, 25] have enabled impressive emergent abilities across many different domains. Resultantly, LLMs are becoming ever more prevalent in scientific literature, as well as in wider society. A developing branch of LLM research seeks to widen the spectrum of achievable tasks by incorporating additional modalities, in particular, the visual modality. Notable multimodal large language models (MLLMs) include PaLM-E [9], Flamingo [3], LLaVA-1.5 [19], InstructBLIP [7], IDEFICS [16], Qwen [4], Kosmos-2 [27], and recently, GPT-4V [26]. These models have shown promising potential demonstrating OCR-free mathematical reasoning [9], diagram and document reasoning [39], and some capacity in medical image interpretation [17, 37]. While there is a growing body of literature examining the geographic and geospatial capabilities of pure language models [22, 23] – including [30], which provides an investigation into the

factual and application-centric geographic capabilities of LLMs – there is yet to be such study of multimodal LLMs.

A comprehensive understanding of the geographic capabilities of MLLMs has the potential for a plethora of downstream societal and research applications. The ability to interpret natural images and maps could aid navigation, routing and localisation, and help answer complex questions. Similarly, understanding remote sensing data – *ex-situ* data derived from air or spaceborne platforms – could empower many research areas, particularly environmental research and disaster response. Conducting an evaluation of the geographic capabilities of MLLMs is especially timely as the current generation of state-of-the-art models – and principally, the recent release of frontier model GPT-4V [26] – have surpassed a critical threshold where they have sufficient capabilities to understand and at least attempt to perform most complex visual tasks.

However, due to the breadth of the abilities of large models, and the nature of emergent abilities [34], it is difficult to know *a priori* the tasks a specific model can perform. A corollary to this is the added complexity of model stochasticity, which makes a comprehensive characterisation of ability more problematic. These difficulties are exacerbated by the closed-source nature of some models and the challenges associated with understanding the distribution of data within a pretraining set on the scale of billions or trillions of tokens. These complexities necessitate exploration by experimentation in a manner that more closely resembles inductive rather than deductive reasoning: deriving general hypotheses and heuristics regarding performance on different tasks based on mostly qualitative observations.

To this end, we devise and carry out a set of visual geographic experiments that probe the key geographic and geospatial knowledge and reasoning abilities of a suite of MLLMs in both qualitative and quantitative settings, with a focus on determining the extent of the abilities of the frontier model, GPT-4V. To aid reproducibility and facilitate future model comparisons and benchmarking, we release a small-scale benchmark of our experiments. This dataset partially contains newly generated samples which were not crawled from the Internet. This strategy helps to mitigate test set contamination, in which models encounter evaluation instances during training. Finally, we distill our findings into the following **key takeaways** to inform the research community:

- Of all the evaluated models, GPT-4V can perform the broadest range of tasks. However, it does not always perform best, *e.g.*, satellite image detection and classification tasks. In general, it recognises fine-detail well but tends to fail when precise localisation is required.
- More broadly, the best model choice depends on the task at hand. Qwen-VL and LLaVA-1.5 in particular often demonstrate good localisation performance.

- Enforcing a specific output format is challenging, models often resort to explanations why they are not capable of performing the task. Among the evaluated models GPT-4V was least susceptible to this behaviour.
- The current generation of leading MLLMs suffer a performance penalty when processing multi-object images, relative to their performance on single object images.

2. Related Work

Geographic Capabilities of LLMs. A comprehensive study of the performance of various LLMs on different geospatial intelligence tasks is given by [22]. Other works investigate more specific tasks, such as the extraction of geospatial knowledge [23] or as a mapping assistant [13]. GPT4GEO [30] includes a mixture of qualitative and quantitative evaluations of the geographic capabilities of GPT-4 across various knowledge-based and reasoning experiments. We take inspiration from [30], building upon it by (1) expanding our experiments to include visual inputs; (2) including comparisons to open-source models; and (3) releasing a small-scale experimental benchmark.

Geographic Capabilities of MLLMs. There is yet to be an extensive evaluation of MLLM capabilities in the geographic domain. However, various geographic or geospatial examples can be found in more general evaluations. In [22], the performance of OpenFlamingo-9B on a remote sensing classification dataset is evaluated. A handful of satellite image descriptions from GPT-4V and Bard are evaluated in [24]. The broad survey conducted in [39] includes a small number of samples of GPT-4V identifying landmarks, foods and interpreting maps. Our work includes all of these tasks to a greater depth as well as introducing many others, including particular attention to the interpretation of remote sensing data, a crucial data source for vision research.

Other Capabilities of MLLMs. Traditional evaluations of MLLMs are carried out as defined benchmarks, such as MME [10], SEED-Bench [18] and MMBench [20], totaling 1000s of questions and corresponding answers. However, the recent release of GPT-4V has sparked a flurry of more qualitative and open-ended evaluations that probe capability beyond a simple question-answer setting. Examples of such works for specific tasks include using GPT-4V for embodied decision-making [6], for self-driving [35] and various medical applications [17, 37]. More general studies, such as [39] includes a broad collection of intriguing general vision examples and [38], which also includes interpretation of infrared and depth imagery, as well as mel spectrograms. We build upon these preliminary analyses of [39] and conduct a detailed exploration focused on the capabilities of GPT-4V in the geographic domain while adding quantitative experiments and comparisons to other MLLMs.

3. Methodology

3.1. Experimental Design

In designing our experiments, we draw inspiration from [30] and create a suite of both qualitative and quantitative experiments. We categorise our experiments according to the type of visual input and include geographic tasks covering natural (photographs), abstract (*e.g.*, maps/flags), and remotely sensed (satellite) images. We adopt this approach of providing a broad capability overview instead of a systematic study, for two main reasons: (1) at the time of writing, the GPT-4V API is rate-limited to 100 requests/day, too few for a large-scale study, (2) an automated evaluation would require structured output, however, current models cannot reliably be controlled for this when pushed close to the limits of their abilities. As LLMs are trained on larger web-crawled datasets, an increasing problem for evaluation is data contamination [21], *i.e.* test data has been seen during training. Although we leverage some established datasets, most of our evaluation is conducted on visual inputs that are sufficiently curated to make their existence in the model training distribution highly unlikely. Consequently, we adopt a methodology that could deliver insights despite these limitations.

3.2. Models

We focus our experimentation on the current most capable model, GPT-4V [26], and query it through both the ChatGPT interface² and API³. As a comparison, we also evaluate the open-source LLaVA-v1.5-13b [19], IDEFICS-80b-Instruct [16], Qwen-VL-Chat [4], InstructBLIP-Vicuna-13b [7] and Kosmos-2 [27] models, accessed predominantly through HuggingFace Transformers [36]. See Appendix for specific model hyperparameters. Hereafter, references to these models refer to the specific versions mentioned here. For some experiments, only a subset of the open-source models could perform the requested task; in these cases, we only evaluate those that could.

3.3. Prompting

Given the wide range and uncharted nature of our experiments, we treat prompting as a hyperparameter that we tune for each model and experiment. We leverage prompt engineering techniques (*e.g.*, 0-shot CoT [15]) where relevant and discuss their impact, however, a systematic comparison of strategies is not appropriate for this setting.

²<https://chat.openai.com/>

³<https://openai.com/blog/openai-api>

4. Experiments

4.1. Localisation

Accurately inferring the location an image is taken has applications ranging from forensic analysis to navigation in GPS-denied environments. To probe this capability, we take inspiration from the popular geography game, *GeoGuessr*⁴, in which players are placed somewhere in the world in a street view panorama and have to guess their location. We construct a dataset of 100 images by randomly sampling a single image for each sovereign state in the GeoLocation dataset [14]. Using the prompt shown in Fig. 1, we query each model to predict the country each image was taken in. We find that GPT-4V in particular is able to extract small details from the images (such as species of foliage, road signs, advertised products, and architecture) and reason over them when making a prediction – see Fig. 1b. In this experiment, we do not detect strong geographic biases from GPT-4V, although countries in S. America were frequently mistaken for other continents (Fig. 1c). Though inferior to GPT-4V, Qwen-VL and InstructBLIP performed well (Fig. 1d); on the other hand, a proclivity to overpredict the USA prevented LLaVA-1.5 from scoring well.

4.2. Remote Sensing

The ability to interpret remotely sensed (RS) data enables numerous positive applications including environmental research and land-use planning. In the following section, we outline a series of tasks involving RS data that test capabilities ranging from holistic image classification and understanding to fine-grained detail extraction and localisation.

Classification. As current MLLMs are limited to pure language output, we initially explore their ability to classify RS imagery with image-level labels. To broadly characterise this ability in the RS domain we evaluate performance on a subset of the SATIN metadataset [29], a challenging benchmark that includes different resolutions, fields of view sizes, class categories, and imagery types. We take a 12-image class-balanced sample from two randomly selected datasets from each of the 6 tasks in the benchmark (totaling 144 images). The zero-shot classification results on this subset are shown in Tab. 1. There is some variation in ranking across the different tasks, though overall LLaVA-1.5 performs best. A mean accuracy of 0.56 for LLaVA-1.5 might seem low; however, this represents the challenge of the benchmark – vision-language baselines *e.g.* CLIP [28] score comparably on the overall SATIN benchmark [29]. A necessary caveat to this comparison is that we only evaluate a narrow subset of the benchmark. To partly address this, we evaluate LLaVA-1.5 on 4 additional subsets of the data (resampled with replacement) and report a low vari-

⁴<https://www.geoguessr.com/>

Zero-Shot Classification Accuracy <i>per task</i>							
<i>Model</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>Mean</i>
GPT-4V	0.50	0.67	0.57	0.53	0.46	0.38	0.52
LLaVA-1.5	0.42±0.2	0.71±0.1	0.50	0.48	0.71±0.1	0.54	0.56
Qwen-VL	0.25	0.63	0.40	0.37	0.50	0.38	0.42

Table 1. **0-shot satellite imagery classification** on a subset of SATIN [29]. For LLaVA-1.5 we report $\mu \pm \sigma$ across 5 subsets.

ance, giving us confidence that the subset used for evaluating the other models is representative.

Season Detection. A key usage of RS data is monitoring and more specifically, *detecting change*. The ability to identify fine-grained differences in an image time series and conduct a nuanced analysis of the cause has many downstream uses. It is not clear the extent to which current MLLMs can perform this task, though preliminary investigations, *e.g.*, [38] suggest they struggle. We conduct a small-scale experiment probing the degree to which GPT-4V can determine the season of images in time series from [33]. We select 6 four-season quadlets (same location) that are clear with minimal atmospheric effects and evaluate GPT-4V’s interpretation. This task is particularly challenging as the imagery contains only minor differences that are difficult to attribute to a particular season, hence GPT-4V scores $\sim 38\%$ accuracy. However, this relatively low accuracy is not due to an inability to detect changes, but rather to their interpretation. Fig. 2 illustrates GPT-4V’s correct analysis of one of the quadlets, demonstrating good detection of fine-grained details and reasoning. We conclude that GPT-4V displays impressive, though fledgling capabilities to interpret fine-grained details in time-series RS imagery.

Segmentation. Having demonstrated the ability to describe fine-grained details in images, we increase the task complexity to segmentation: identifying *and* localising objects and classes. Despite not being able to provide visual outputs, it is possible – with creative prompting – to create segmentation maps from pure language. We introduce two such prompting strategies: (1) **Grid Segmentation:** Segment the image into the following [classes]. Display the results as a [YxY] table with each cell labelled with one of the class labels, and (2) **SVG Segmentation:** Segment the image into the following [classes]. Provide the code for an SVG that displays the segmentation map. We include examples of each prompting strategy in Fig. 3. Clearly, the segmentation maps are not perfect and underperform vision-language or supervised vision baselines; however, several positive observations can be made. Generally, the relevant classes are detected and included in the map. Additionally, key features such as building groups, roads, and water bodies are detected and localised in approximate positions. Furthermore, annotation inaccuracies are demonstrated as the

Mean Distance Error (km)							
Projection	North				South		Mean
	<i>Africa</i>	<i>Asia</i>	<i>Europe</i>	<i>America</i>	<i>Oceania</i>	<i>America</i>	
Mercator	3542	2486	1351	2809	1657	3460	2551
Miller	9398	3391	990	3334	2071	7434	4436
Mollweide	8441	2985	1289	2587	2798	2998	3516
PlateCarree	9339	2730	1874	3032	2058	3215	3708
Robinson	2796	2501	1040	3661	1984	3230	2535
Mean	6703	2819	1309	3084	2114	4067	3349

Table 2. **Localisation: map \rightarrow real-world.** The mean distance error is calculated over 10 points on each continent map crop.

ground truth is mostly labelled as background but perhaps more appropriately given the forest label suggested by GPT-4V. Considering a broader range of examples, we observe grid segmentation to outperform SVG segmentation. We find that there is not sufficient capacity in the other MLLMs to perform segmentation.

Bounding Boxes. We also analyse the object detection capabilities of the MLLMs in the form of bounding boxes. We draw a number of qualitative observations from the example bounding boxes in Fig. 4. As reported in [39], we find that bounding boxes provided by GPT-4V are mostly inaccurate. Qwen-VL produces tight, accurate boxes that are more accurate than LLaVA-1.5. Kosmos-2 also produces accurate bounding boxes and demonstrates a different – though arguably also correct – interpretation of ‘urban area’. The bounding box from IDEFICS is especially poor.

Counting. Counting small objects in images probes the resolution of the visual component of each MLLM. In Fig. 5, we include an aerial scene from [8] comprising 45 sea lions of various sizes lounging on a wave-beaten rock of similar colouration to the sea lions. This particular task is difficult – GPT-4V’s prediction suggests it is close to the limits of what it can perceive, while it is clearly beyond the capabilities of the other models.

4.3. Mapping

Interpreting maps is a core geospatial skill required for route planning, urban development, and disaster response. The fundamental requirement for successful map reading is relating areas and positions on maps to the real-world locations they represent. To evaluate this, we carry out experiments that focus on identifying geographic entities from maps, localising points on maps, and map annotation.

Region Identification. We collect a compact dataset of three identification tasks: (A) state from outline shape, (B) island or water body naming, and (C) city from map. Each task comprises 16 samples generated by the authors, minimising the risk of test-set contamination. These experi-



Figure 2. **Satellite imagery season detection.** We test GPT-4V’s ability to detect seasonal changes in a four-image time-series from [33]. In this example, the model is able to pick up minor details such as crop colouration and the presence of snow to correctly estimate seasons.

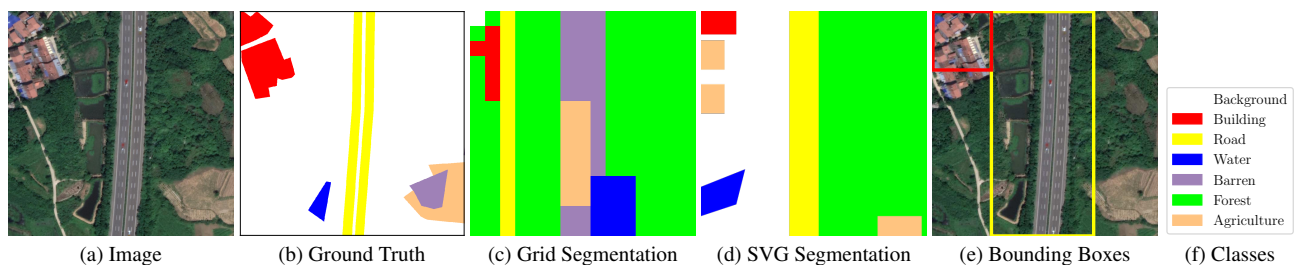


Figure 3. **Segmentation using GPT-4V.** We include examples of Grid (c) and SVG (d) segmentation, and localisation (e) of satellite imagery (a) from LoveDA [32]. Bounding boxes are for **urban areas** and **road**. Segmentation labels are given in (f).

ments assess pattern and shape processing capabilities and, visual world knowledge. We show the results in Fig. 6.

A. State name from outline. In the first experiment, the models are tasked with inferring a state’s name from the shape of its outline. To obtain data, we use GeoPandas [12] to render state outlines. The results show a good performance of GPT-4V and Qwen-VL while InstructBLIP and LLaVA-1.5 perform worse.

B. City name from maps. Given a city map image, we prompt the models to identify the corresponding city. To generate the data, we use the OpenStreetMap-based Maputnik tool [1] and extract maps at zoom-level 12 at the city locations provided by the GeoNames database [11]. Cities are sample-balanced over the continents. We extract images of size 1000×1000 pixels. The responses of GPT-4V suggest it captures the patterns in the map data but lacks the visual knowledge to associate them with specific locations on Earth. The other models tend to perform worse. LLaVA-1.5 (and GPT-4V to a smaller degree) often predict well-known cities like New York or San Francisco indicating a bias towards presumably frequently mentioned cities in the training data. These cities also count among the few samples the models classified correctly, the models were consistent in this. GPT-4V describes scenes well but appears to fail to

retrieve the right name. This could be because information encoded in city maps is rarely expressed in text and only selected city maps were seen during training. InstructBLIP often outputs nonsensical answers for the provided prompt, highlighting a deficit in instruction following.

C. Island and water body naming from maps. In the last experiment, islands and water bodies need to be identified. We use the same method for obtaining data as in the previous city map experiment (B) but select locations and zoom level manually. Furthermore, we simplify the maps to only incorporate land cover and sea, since we found this to work better in early experiments. The task is easier than understanding city maps but still challenging as even the best model, GPT-4V, correctly classified only 6 out of 16 samples.

Localisation: map \rightarrow real-world. Next, we test the accuracy of the models in predicting the real-world positions (latitude and longitude) of query points on maps. To do so, we create equal latitude and longitude extent map crops (using Cartopy [2]) of each continent and place ten coloured points on each. We repeat this process to create equivalent maps for 5 common map projections, and prompt GPT-4V with: ‘*Estimate the Latitude/Longitude positions of the coloured points on the map*’, to obtain a position estimate

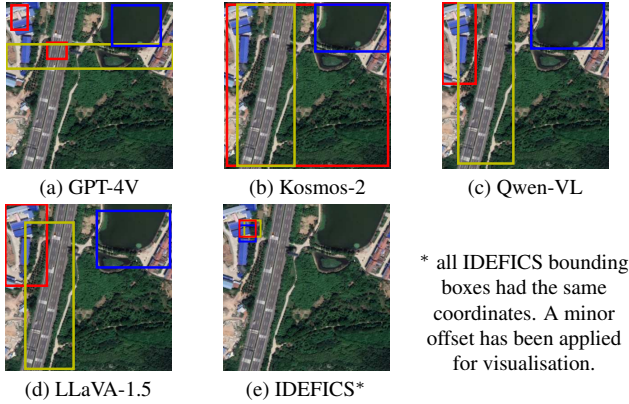


Figure 4. Bounding boxes for urban areas, road and water bodies.

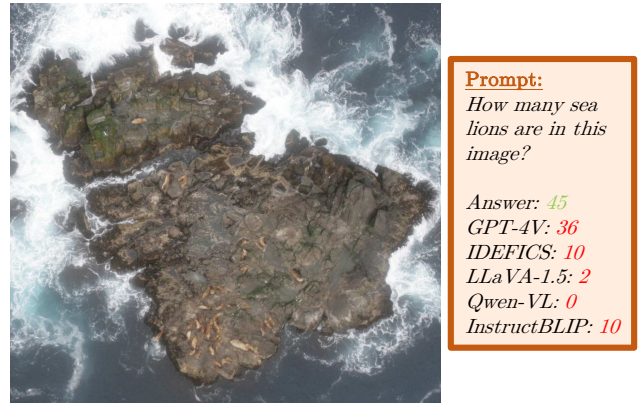


Figure 5. Counting small objects proves challenging.

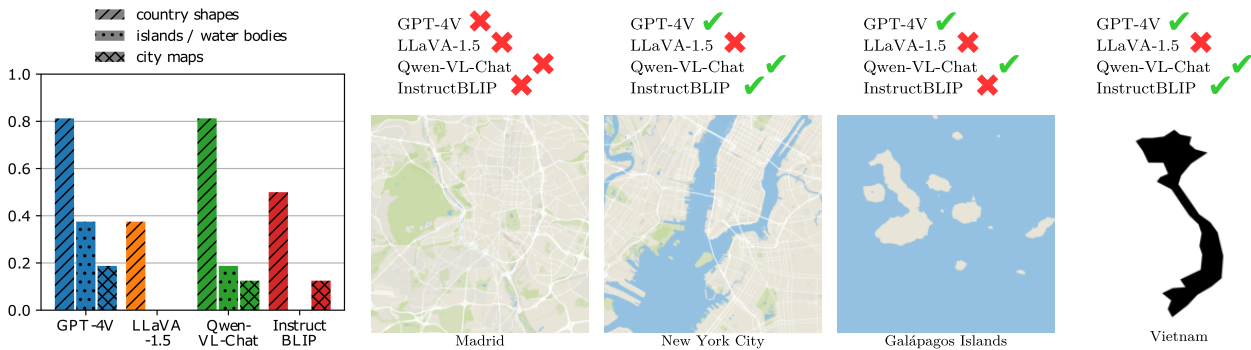


Figure 6. Region identification. (a) Quantitative results. (b) Example images (note, high-resolution images were provided to the models.)

for each point. Using the haversine formula, we calculate the error between the true (circles) and predicted (crosses) positions of each point (Fig. 7); the mean position errors for each map are shown in Tab. 2. Several insights can be gleaned from these results, firstly that GPT-4V struggles with the task, scoring an average error of nearly 1000 km on its most accurate configuration (Europe, Miller). Frequently, GPT-4V confuses the continent the points are placed on, such as predicting positions in the Americas for points in Africa (Fig 7). A geographic bias is evident with the lowest errors in Europe and significantly higher errors in Africa. This is in agreement with other results, such as the difficulties identifying African countries (see *Failure cases*), and plotting Africa’s outline in [30], suggesting that geographic information about Africa is less prevalent in the training data. We also observe variation in error amongst the different projections, with the Mercator and Robinson proving the easiest for GPT-4V to interpret. This variation could be caused by factors such as the interpretability of specific projections (especially at extreme latitudes) or their prevalence in the training distribution.

Localisation: real-world → map. The final component of the cartographic experimentation tests the ability to-

calise in the inverse direction: annotating points on the map given a real-world location. Concretely, given the names of European capital cities, we prompt GPT-4V to provide their positions on a bespoke coordinate system overlaid on a map of Europe (created using Cartopy) – details of the inputs and prompts are illustrated in Fig. 9. We analyse GPT-4V’s performance on two different task settings: (1) **Image** in which we pass the map and grid to GPT-4V and (2) **No Image** in which we just pass a brief description of the map. When shown the map, GPT-4V struggles to accurately locate the capital cities on the grid, despite the grid cells being relatively coarse, locating only 1 capital within the correct grid cell. Conversely, when simply prompted with a description of the map – including the *extent* – GPT-4V is able to locate *every* capital in the correct grid cell. It does so by performing the following: (i) estimate the approximate latitude/longitude of each capital using its knowledge, (ii) determine a transformation from latitude/longitude to the grid reference, and (iii) perform the conversion. This dichotomy in performance highlights GPT-4V’s strong capabilities to solve complex problems in a pure language setting while struggling to perform precise and accurate image analysis. Performance varies depending on the grid resolution.

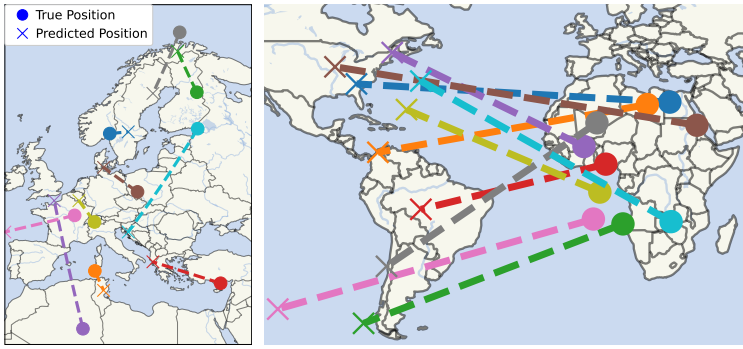


Figure 7. **Localisation: map → real-world:** predicted latitude and longitude positions. [left] Europe (Miller), [right] Africa (PlateCarree).

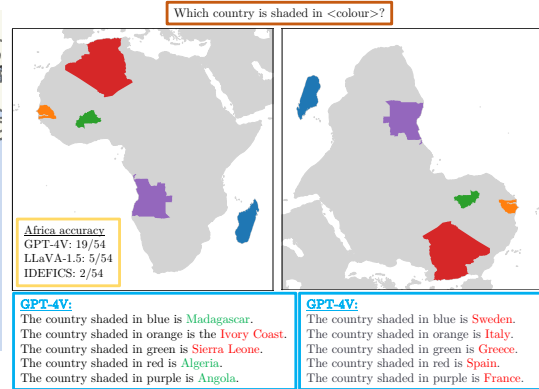


Figure 8. **Identifying multiple states** proves difficult on a regular map and is a clear failure case when the map is flipped.

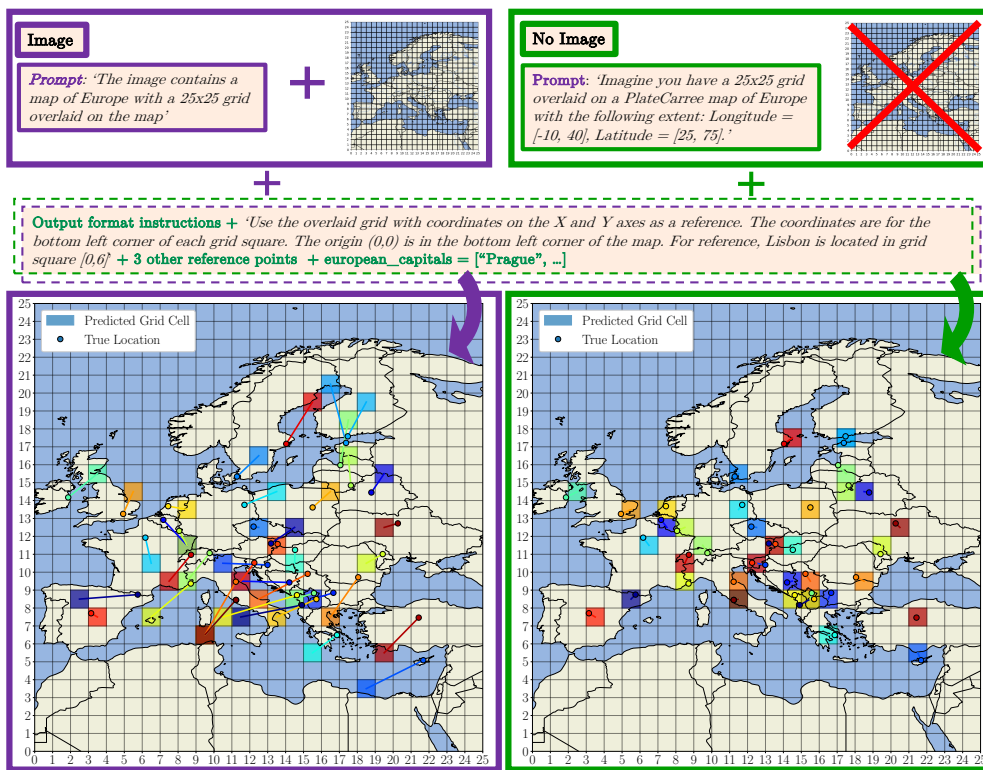


Figure 9. **Localisation: real-world → map.** Using GPT-4V to annotate European capital cities on a map overlaid with a grid in two settings: **Image** [left] where the map grid is passed to the model and **No Image** [right] where just a brief description of the map is passed.

4.4. Flags

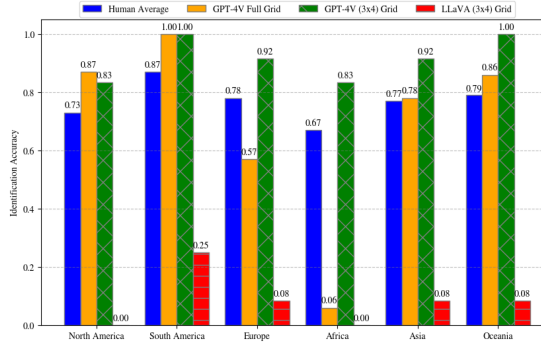
Identifying multiple flags in images presents a unique challenge that balances the visual and knowledge-based aspects of MLLMs. This is particularly relevant to understanding the geographic biases inherent in these models and their capacity to handle different scales of visual data. Our experiments employed flag images sourced from Sporcle quizzes⁵ that were divided by continent, cropped and presented to the models in grids of varying dimensions, see Fig. 10a. The

⁵<https://www.sporcle.com/>

analysis of the comparative identification performance data (Fig. 10b) reveals a discernible geographic bias in model performance, with GPT4-V excelling in North and South America. In contrast, its performance on African flags shows a pronounced -61.0% accuracy differential compared to human benchmarks. While this may suggest a potential under-representation of African nations in the training data, it is crucial to consider that other factors may contribute to this outcome, such as grid size, image resolution, and quality. GPT-4V's proficiency is apparent in its consistent



(a) Example flag identification input & GPT-4V response.



(b) Model comparison on flag identification accuracy.

Figure 10. **Flag identification.** (a) Presents the input used in the full grid case, with country names beneath added after to indicate correct and incorrect predictions. (b) Depicts prediction accuracy of different models, with human averages from Sporcle.

identification accuracy across the standardized 3x4 grids, whereas other models falter, often giving non-sensical answers to larger grid formats.

4.5. Failure Cases

Our analysis focuses on tasks where there is sufficient capability in the models to attain reasonable performance. However, we encountered many, at times surprising, failure cases. Settings in which GPT-4V struggled include routing using maps, drawing and improving country outlines, annotating missing labels on travel maps, estimating population growth from satellite time series, and determining the elevation profile of mountains. Additional details and examples of these cases are included in the Appendix.

Multiple States. Using Cartopy, we create a set of maps of Africa with 5 countries randomly shaded in different colours (Fig 8). This task represents a variation of the *State name from outline* experiment, in which additional context is given but there is the added complexity of identifying multiple regions from a single image. We find this task proves challenging: GPT-4V scores $\sim 35\%$ while LLaVA-1.5 and IDEFICS correctly identify only 5 and 2 African states, respectively. The models are completely unable to correctly identify states when the map is rotated by 180° .

5. Conclusions

We evaluate a selection of state-of-the-art MLLMs on a suite of experiments that explore various vision capabilities in the geographic and geospatial domains. By creating a large proportion of our test data we reduce test contamination, avoiding memorisation. We distill our analysis into key takeaways to clearly inform the research community of MLLM capabilities in this domain. We find that the current generation of MLLMs possesses remarkable abilities to interpret geographic visual data, though struggle with a number of tasks, such as route planning with maps and abstract navigation. Overall, we find GPT-4V to out-

perform the other models in most settings, and its strong instruction-following ability enables it to attempt a much broader range of tasks. We demonstrate that GPT-4V is able to extract fine-grained details from imagery but is weak at localising and struggles to draw accurate bounding boxes. Of the open-source MLLMs we evaluate, LLaVA-1.5 and Qwen-VL prove to be the most capable and are better suited to some tasks, *e.g.* object localisation, than GPT-4V. We also introduce two segmentation prompting strategies and release our experimental data as a small-scale geographic benchmark for future evaluation.

Limitations. Where possible we maximise the robustness of our experimentation with repeats, consistent prompting and model hyperparameters, normalising inputs and comparisons to ground truth data. However, this robustness is inherently limited by the accessibility of models, specifically API access limits that prevent large-scale quantitative experimentation. Future work could include quantitative evaluation with multiple-choice style questions.

Broader Impacts. We demonstrate the capabilities of MLLMs to detect fine-grained details in images and perform nuanced reasoning, with potential applications in environmental research and disaster response. We also highlight numerous cases where the current models are severely lacking, especially in map interpretation. Our analysis suggests there are geographical biases in the data, with weaker performance consistently shown for regions such as Africa that are perhaps less represented in the training distribution.

Acknowledgements. This work was supported by the UKRI CDT in Application of AI to the study of Environmental Risks (ref. EP/S022961/1), an Isaac Newton Trust grant, the Hong Kong Research Grant Council Early Career Scheme (Grant No. 27208022), the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project-ID 494541002, and an EPSRC HPC grant. SA would like to acknowledge the support of Z. Novak and N. Novak in enabling his contribution.

References

- [1] Maputnik: A free and open visual editor for the maplibre gl styles. <https://maputnik.github.io/editor/>. Accessed: 2017-11-15. 5
- [2] CartoPy: A python library for cartographic visualizations with matplotlib. <https://scitools.org.uk/cartopy/docs/latest/>, 2023. Accessed: yyyy-mm-dd. 5
- [3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 1
- [4] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report, 2023. 1, 3
- [5] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrike, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023. 1
- [6] Liang Chen, Yichi Zhang, Shuhuai Ren, Haozhe Zhao, Zefan Cai, Yuchi Wang, Peiyi Wang, Tianyu Liu, and Baobao Chang. Towards end-to-end embodied decision making via multi-modal large language model: Explorations with gpt4-vision and beyond. *arXiv preprint arXiv:2310.02071*, 2023. 2
- [7] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 1, 3
- [8] Meg Risdal DataCanary, Katie. Noaa fisheries steller sea lion population count, v1. <https://www.kaggle.com/competitions/noaa-fisheries-steller-sea-lion-population-count/data>, 2017. 4
- [9] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palme: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 1
- [10] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 2
- [11] GeoNames. Geonames. <https://www.geonames.org>. 5
- [12] Kelsey Jordahl, Joris Van den Bossche, Jacob Wasserman, James McBride, Jeffrey Gerard, Martin Fleischmann, Jeff Tratner, Matthew Perry, Carson Farmer, Geir Arne Hjelle, et al. geopandas/geopandas: v0. 6.0. *Zenodo*, 2019. 5
- [13] Levente Juhász, Peter Mooney, Hartwig H Hochmair, and Boyuan Guan. Chatgpt as a mapping assistant: A novel method to enrich maps with generative ai and content derived from street-level photographs. *arXiv preprint arXiv:2306.03204*, 2023. 2
- [14] Rohan K. Geolocation - geoguessr images (50k), v1. <https://www.kaggle.com/datasets/ubitquitin/geolocation-geoguessr-images-50k/>, 2022. 3
- [15] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023. 3
- [16] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents, 2023. 1, 3
- [17] Seowoo Lee, Jiwon Youn, Mansu Kim, and Soon Ho Yoon. Cxr-llava: Multimodal large language model for interpreting chest x-ray images, 2023. 1, 2
- [18] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 2
- [19] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 1, 3
- [20] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 2
- [21] Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Dublin, Ireland, 2022. Association for Computational Linguistics. 3
- [22] 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. 1, 2
- [23] Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David Lobell, and Stefano Ermon. Geollm: Extracting geospatial knowledge from large language models. *arXiv preprint arXiv:2310.06213*, 2023. 1, 2
- [24] David Noever and Samantha Elizabeth Miller Noever. Multimodal analysis of google bard and gpt-vision: Experiments in visual reasoning, 2023. 2
- [25] OpenAI. Gpt-4 technical report, 2023. 1

- [26] OpenAI. Gpt-4v(ision) system card. 2023. [1](#), [2](#), [3](#)
- [27] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. [1](#), [3](#)
- [28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. [3](#)
- [29] Jonathan Roberts, Kai Han, and Samuel Albanie. Satin: A multi-task metadataset for classifying satellite imagery using vision-language models. *arXiv preprint arXiv:2304.11619*, 2023. [3](#), [4](#)
- [30] Jonathan Roberts, Timo Lüddecke, Sowmen Das, Kai Han, and Samuel Albanie. Gpt4geo: How a language model sees the world’s geography. *arXiv preprint arXiv:2306.00020*, 2023. [1](#), [2](#), [3](#), [6](#)
- [31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. [1](#)
- [32] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation, 2022. [5](#)
- [33] Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M Albrecht, and Xiao Xiang Zhu. Ssl4eo-s12: A large-scale multi-modal, multi-temporal dataset for self-supervised learning in earth observation, 2023. [4](#), [5](#)
- [34] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022. [2](#)
- [35] Licheng Wen, Xuemeng Yang, Daocheng Fu, Xiaofeng Wang, Pinlong Cai, Xin Li, Tao Ma, Yingxuan Li, Linran Xu, Dengke Shang, Zheng Zhu, Shaoyan Sun, Yeqi Bai, Xinyu Cai, Min Dou, Shuanglu Hu, and Botian Shi. On the road with gpt-4v(ision): Early explorations of visual-language model on autonomous driving, 2023. [2](#)
- [36] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface’s transformers: State-of-the-art natural language processing, 2020. [3](#)
- [37] Chaoyi Wu, Jiayu Lei, Qiaoyu Zheng, Weike Zhao, Weixiong Lin, Xiaoman Zhang, Xiao Zhou, Ziheng Zhao, Ya Zhang, Yanfeng Wang, and Weidi Xie. Can gpt-4v(ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis, 2023. [1](#), [2](#)
- [38] Yang Wu, Shilong Wang, Hao Yang, Tian Zheng, Hongbo Zhang, Yanyan Zhao, and Bing Qin. An early evaluation of gpt-4v (ision). *arXiv preprint arXiv:2310.16534*, 2023. [2](#), [4](#)
- [39] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of llms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 2023. [1](#), [2](#), [4](#)