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

Supplementary Material

We structure this Appendix to our main paper in two parts: **(1)** we provide examples of failure cases in which the models, including GPT-4V, were unable to adequately perform a given task, and **(2)** we outline specific details of the experiments we discuss, including information regarding hyperparameters and prompts.

1. Failure cases

1.1. Identifying multiple states

Having created the map images with multiple shaded countries in Africa as mentioned in the main paper, we attain the results for the accuracy of identifying multiple states simultaneously based on the following experimental procedure. We pass each map to the models (GPT-4V, LLaVA-1.5, IDEFICS-80b-Instruct) with the following prompt:

```
Using the map given above, answer the
following:  Which country is shaded in blue?
Which country is shaded in orange?  Which
country is shaded in green?  Which country
is shaded in red?  Which country is shaded in
purple?
```

We query GPT-4V using the ChatGPT interface and run LLaVA-1.5 and IDEFICS using top_p set to 0.7 and temperatures of 0.0 and 0.01, respectively. The results we report in the main paper (GPT-4V scoring 19/54, LLaVA-1.5 scoring 5/54 and IDEFICS scoring 2/54) indicate that the task is challenging. The weaker models that score close to 0 are unable to interpret and localise the countries in the image correctly, though in their response descriptions and predictions did mostly correctly identify the continent as Africa. GPT-4V scored well above the chance score but favoured larger states in its predictions and correct answers and rarely correctly identified small states, especially in regions with multiple small states.

1.2. Map routing

We explored a variety of different experimental variations involving interpreting journeys using maps.

In the first setting, we provided GPT-4V with a street map annotated with a start and end marker and asked for a description of a route between the two using the following prompt:

```
Describe the route to walk from the location
marked with the {start_marker} to the location
marked with the {end_marker}.
```

However, we found that in every instance (~25 samples) this resulted in a route that was incorrect and involved hallucinations, such as this response to the annotated OpenStreetMap (OSM)[1] map shown in Fig. 1:

```
1. Start at the red circle located on
   Mawson Road.
2. Head northeast on Mawson Road
   toward St.  Philip's Road.
3. Turn right onto St.  Philip's Road.
4. Continue on St.  Philip's Road
   until you reach Mill Road.
5. Cross Mill Road and continue
   straight onto Romsey Road.
6. Follow Romsey Road as it curves to
   the left.
7. Continue on Romsey Road until you
   reach the green circle near the
   junction with Vinery Road.
```

We repeated these experiments across different locations, marker colours and shapes, and map styles, however, incorrect routes were always suggested.

Another setting that we tried involved providing a series of street-level images along a route followed by a map with a start point and start direction annotated. We then prompted GPT-4V to describe where on the map the final destination was. As in the first setting, the final positions described were inaccurate.

Finally, we investigated a simple setting consisting of a grid with one cell marked as the start and another as the finish. After passing a description of the setup and instructions to navigate the grid, as well as some examples, we prompted GPT-4V to describe a journey from the start to finish points. Again, we found that GPT-4V was unable to perform this task correctly.

1.3. Outlines

In [3], qualitative experiments showed that GPT-4 could be prompted to provide reasonably accurate coordinates for the outlines of countries, rivers, etc. For Australia, the authors demonstrated an iterative improvement method in which an initial set of coordinates for the outline of Australia were generated from GPT-4. Then, text feedback was provided to GPT-4 in an attempt to prompt it to provide an improved and more accurate outline. We take inspiration from this

[1] https://www.openstreetmap.org/

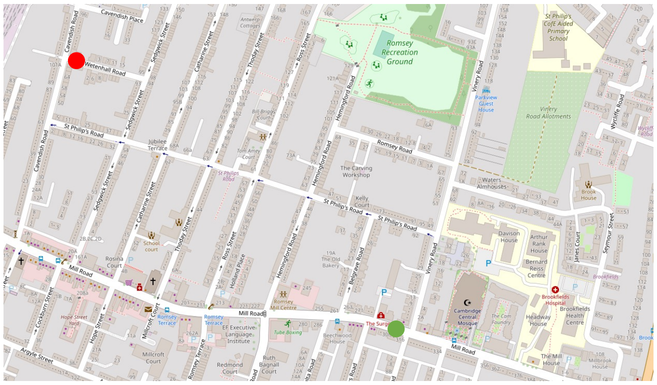


Figure 1. **Example annotated map input.** An example input OSM map of Cambridge annotated with a journey start (red circle) and end point (green circle).

experiment however, rather than providing text feedback we plot the provided outlines on a map and pass it to GPT-4V.

We initially query GPT-4V to provide coordinates for the outlines of specified countries, using this text-only prompt:

```
Provide the latitude and longitude coordinates
for an outline of {country_name}. Use
approximately 50 points, ensuring that the
outline does not overlap itself. Provide
the coordinates as a python list in the
following format: coordinates = [(LAT_COORD1,
LON_COORD1), (LAT_COORD2, LON_COORD2),...]
```

Then we begin an iterative process of providing GPT-4V with the provided outlines overlaid on the map image along with this prompt:

```
I have plotted this outline (see {line_colour}
line). Provide an updated version of
the coordinates given the errors in the
{line_colour} outline.
```

Example country outlines are shown in Fig. 2, illustrating the results of this process. We find that GPT-4V struggles with this task and is unable to consistently improve country outlines. Generally, subsequent sets of coordinate outlines are *less* accurate. We briefly experiment with additional visual markers to aid performance, however, we do not see a noticeable improvement.

1.4. Networks

We investigate the ability of GPT-4V to correctly annotate travel network maps. We modify existing network maps by adding visual prompts such as arrows (Fig. 3)[2]. Next, we pass these images to GPT-4V along with a prompt such as:

```
The image represents a transport network
map for Paris. 3 arrows (purple, blue and
green) have been added to the map pointing at
specific stations represented by dots. Which
station is each arrow pointing to?
```

After trying a number of different network maps with different visual prompts and settings (*e.g.*, all but the target stations annotated) this task proved beyond the reaches of GPT-4V. Generally, the model would be unable to provide station predictions, and when it did the answers were consistently incorrect. Given the scale of travel networks, accurate interpretation requires being able to resolve small details.

1.5. Population estimation

We conducted a population estimation experiment that aimed to assess GPT-4V's ability to predict population growth from satellite imagery. GPT-4V was presented with Google Earth Timelapse[3] images of Lahore from 1989 to 2021 and tasked to estimate population changes without external data. The initial population was given, and GPT-4V was prompted to estimate subsequent years' populations based on visual growth cues, as depicted in Fig. 4a.

The experiment did not succeed as intended. The GPT-4V's population estimates, even with visual data, fell significantly short of the actual figures, as shown in Fig. 4b. Although the model with vision performed better than the base model (GPT-4 with just text prompt), it still could not approach the ground truth. This highlights the complexity of the task, which would be challenging even for expert human analysts.

1.6. Elevation

We asked GPT-4V which mountain range is shown in the image in Fig. 5 through the prompt:

```
Which mountains are depicted in this height
map?
```

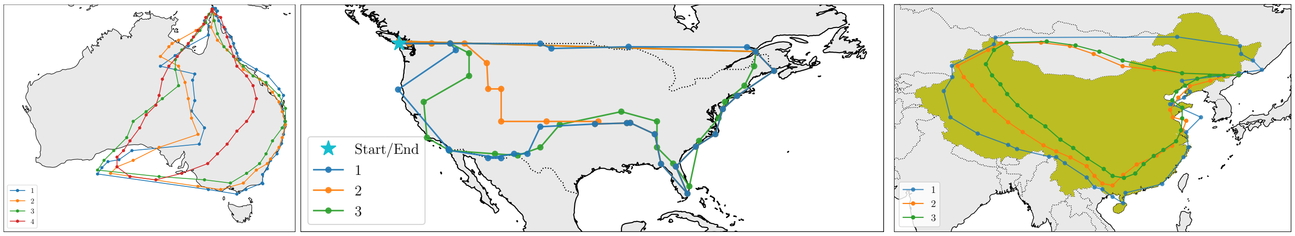
Despite first refusing to answer, after being explicitly asked to make a guess GPT-4V mentioned the following (the actual answer is longer) *The Appalachian Mountains in North America, the Alps in Europe and the Andes in South America*. We consider this a failure case as the model mentioned the correct answer (Alps in Europe) as an option among many incorrect answers.

2. Experimental details

Aside from cases where we interact with GPT-4V via the ChatGPT interface (and hence have no control over the model hyperparameters), we set the value of top_p to 0.7 (0.8 for Qwen-VL) for all experiments and models. We use different values for the temperature parameter depending on

[2] https://commons.wikimedia.org/wiki/File:Paris_Metro_map.svg

[3] https://earthengine.google.com/timelapse/



(a) Australia. (b) USA. (c) China (shaded).

Figure 2. **Country outlines.** We initially query GPT-4V to provide coordinates for the outlines of specified countries. Then we begin an iterative process of prompting GPT-4V with the provided outlines overlaid on the map. We experiment with different visual markers such as specifying the start and finish point (b) or shading the country boundary (c).

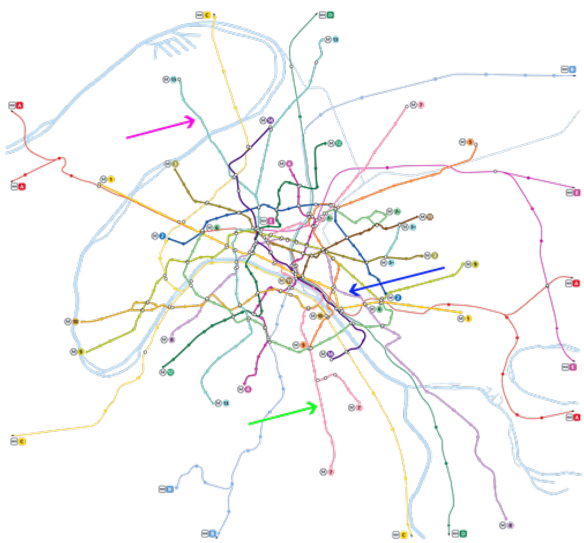


Figure 3. **Network map labelling**. GPT-4V input: Paris transport network. A high resolution image was provided to GPT-4V.

the experiment, the specific values used for each are detailed below.

2.1. Localisation

Preprocessing. We take the GeoLocation dataset [1] and randomly sample one image per sovereign state (other states are dropped). Each image is preprocessed to remove potentially confusing watermarks (*e.g.*, overlaid map). Rather than completely removing all watermarks, which will significantly decrease the image size and field of view, we instead take a crop removing the rightmost 284 pixels.

Prompts. For the localisation (GeoGuessr) experiment, we pass the models a single image along with the following prompts. Where relevant, model temperature is included in parenthesis.

GPT-4V

```
Which country is this picture taken in?  If
you are not certain then provide an educated
guess of a specific country.  Let's think step
by step.
```

LLaVA-1.5 (*temperature*=0.001)

```
Which country is this picture taken in?  If
you are not certain then provide an educated
guess of a specific country.  Let's think
step by step.  You must provide a predicted
country.
```

Qwen-VL (*temperature*=0.1)

```
Question:  Which country is this picture taken
in?  If you are not certain then provide an
educated guess of a specific country.  Let's
think step by step.  You MUST provide a single
predicted country.  Answer:
```

followed by (if an answer is not provided)

```
Pick a single most likely country.  Make an
educated guess if you are unsure.
```

then

```
Pick a single most likely country.
```

then

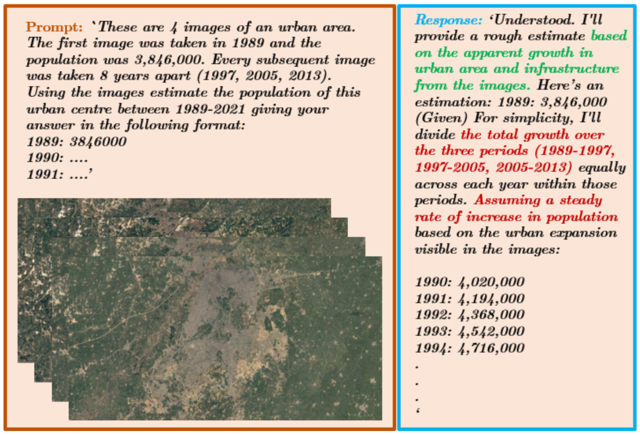
```
Pick the most likely continent.
```

InstructBLIP (*temperature*=0.1)

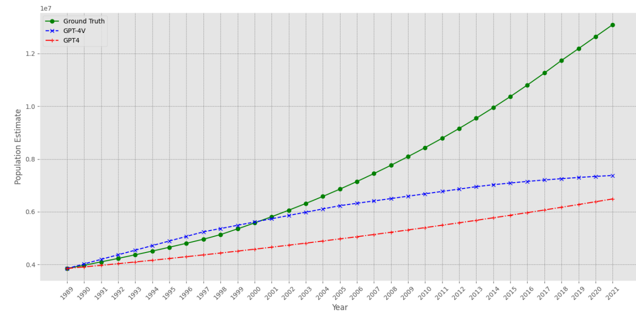
```
Which country is this picture taken in?  If
you are not certain then provide an educated
guess of a specific country.  Let's think
step by step.  You must provide a predicted
country.
```

2.2. Remote sensing – classification

We randomly select the following two datasets from each task in the SATIN metadataset [2]: NASC-TG2, SAT-6 (Task 1), WHU-RS19, EuroSAT (Task 2), RSI-CB256, Million-AID (Task 3), MultiScene, AID-MultiLabel (Task 4), Post Hurricane, UTSC_SmokeRS (Task 5), and Brazilian Coffee Scenes, Brazilian Cerrado-Savanna Scenes (Task



(a) Prompt and response showcasing the model's visual data interpretation ability.



(b) Comparison of population estimates from the GPT-4 and GPT-4V models with the ground truth.

Figure 4. **Population estimation** from satellite imagery task results showing the prompt and model's response (a), and the comparison of the model's estimates to the ground truth (b).

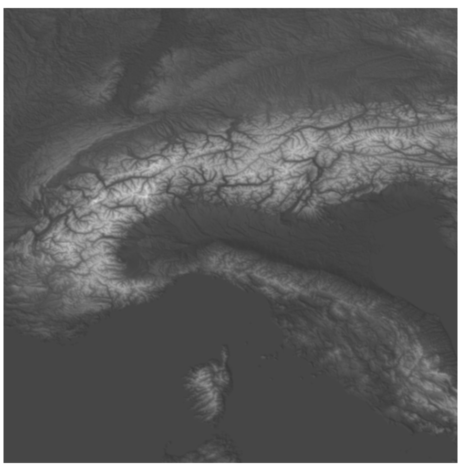


Figure 5. **Elevation**—Height map of the Alps in Europe.

6). For each task, we take a class-balanced subset of 12 images. Images were passed individually to each model with the exception of GPT-4V, in which images were passed in batches of 4 via the ChatGPT interface.

The following prompts were used (where {class_labels} is a comma-separated sequence of the target classes for the dataset):

GPT-4V: Tasks 1-3,5-6

```
Classify each of these images with a single
label from this set: {class_labels}. You
must provide a predicted label for each image.
```

for Task 4

```
Classify each of these images with labels from
this set: {class_labels}.
```

LLaVA-1.5 (*temperature*=0.7): Tasks 1-3,5-6

```
Classify the image with a single label from
this set: {class_labels}. You must provide a
single label.
```

for Task 4

```
Classify the image with labels from this set:
{class_labels}. You must only use labels from
this set.
```

Qwen-VL (*temperature*=0.75): Tasks 1-3,5-6

```
Question: Classify the image with a single
label from this set: {class_labels}. You
must provide a single label. If you are
unsure make an educated guess. Label:
```

for Task 4

```
Question: Classify the image with labels from
this set: {class_labels}. You must provide
predicted labels for the image. Answer:
```

2.3. Remote sensing – change detection

The 4 time series images shown in the main paper were passed to GPT-4V in a single batch along with the prompt shown in the main paper.

2.4. Remote sensing – segmentation

The prompts used for the segmentation example are as follows, using the target land cover classes from the LoveDA dataset [4].

Grid segementation:

```
Segment the image into the following 7 land
cover classes: background (1), building (2),
road (3), water (4), barren (5), forest (6),
agriculture (7). Display the results as a
15x15 table with each cell labelled with one
of the 7 class labels. Don't include any
column or row labels.
```

SVG segmentation:

```
Segment the image into the following 7 land
cover classes: background (1), building (2),
road (3), water (4), barren (5), forest (6),
agriculture (7). Provide the code for an SVG
that displays the segmentation map.
```

2.5. Remote sensing – bounding boxes

We derived bounding box coordinates using the following prompts:
GPT-4V

```
Please follow the instructions 1. Tell me
the size of the input image; 2. Localize each
{object_class} in the image using a bounding
box.
```

LLaVA-1.5, IDEFICS, Qwen

```
Localize the {object_class} in the image using
a bounding box.
```

InstructBLIP

```
Localize the {object_class} in the image using
a bounding box. Bounding box coordinates:
```

Kosmos-2

```
<grounding><phrase> {object_class} </phrase>
```

2.6. Remote sensing – counting

We used the same prompt given in the main paper for each model.

2.7. Mapping – region identification

City Maps We based our analysis on the following 16 cities:
Madrid, Naples, Cairo, Lagos, Buenos Aires, Rio de Janeiro, Mexico City, New York City, San Francisco, Shanghai, Taipei, Mumbai, Tokyo, Stockholm, Cape Town, Vladivostok.

For GPT-4V, we sequentially fed single images into a conversation, which we started by this prompt:

```
I'll show you the map of a city and you tell
me its name. Make only a single guess. The
top of the image is facing north.
```

For the other models we independently asked about single examples using the same prompt (for LLaVA-1.5) or the following slightly amended prompt for Qwen and InstructBLIP:

```
Question: Guess the city that is shown in the
map. Make only a single guess. The top of
the image is facing north. City:
```

Islands and Water Bodies We based our analysis on the following 16 water bodies and islands:
Strait of Gibraltar, Balearic Islands, Sicily, Strait of Hormuz, Bahamas, Galapagos, Samborombón Bay, Falkland Islands, South Georgia, Tasmania, Hawaiian Islands, Öresund, Strait of Malacca, Tahiti, Spitsbergen, Bering Strait.
The selection of places and the exact map crop (including zoom level) were carried out manually.

For GPT-4V, we sequentially fed single images into a conversation, which we started by this prompt:

```
I'll show you the shape of an island or water
body and you tell me its name. Make only a
single guess. The top of the image is facing
north.
```

For the other models we independently asked about single examples using the same prompt (for LLaVA-1.5) or the following slightly amended prompt for Qwen and InstructBLIP:

```
Question: Guess the island or water body
shown in the image. Make only a single
guess. The top of the image is facing north.
Island/Water body:
```

Country Shapes We used the following 16 states for this experiment:
Indonesia, Jordan, Vietnam, Poland, Russia, Belgium, Haiti, Mexico, United States of America, Ethiopia, Tanzania, Malawi, Australia, Colombia, Venezuela, Peru.
They were selected randomly from all countries with a population above 10M while balancing over continents (3 states per continent, except oceania with 1).

For GPT-4V, batches of four images were provided at the same time together with this prompt:

```
Guess the country by the provided shape.
Output nothing but a single guess for each
shape.
```

For the other models we independently asked about single examples using the same prompt (for LLaVA-1.5) or the following slightly amended prompt for Qwen and InstructBLIP:

```
Question: Guess the country by the provided
shape. Output nothing but a single guess for
each shape. Country:
```

2.8. Mapping – Localisation: map $\rightarrow$ real-world

The continent maps were created by taking a crop of 50 degrees latitude and longitude extent. (Note, in Fig. **??** the map showing the predicted positions of a set of test points in Africa has been enlarged beyond the extend of the input image in order to include the predicted positions.) To position the coloured test points we employed stratified random sampling within the map extent with a margin of 2 degrees from the edges (to avoid overlap with the map edge). For a given projection, we queried GPT-4V via the ChatGPT interface in a single chat window using the following prompt on two batches of 3 images (each representing a different continent). Batch one included Africa, Asia and Europe, while Batch two included North America, Oceania and South America.

```
For each of the images, please estimate
the Latitude/Longitude positions of the
coloured points on the map to 2 decimal
places.  A point is included on the map
for each of the following colours so
provide a position estimate for each.
Provide the coordinates as a list in the
following format:  estimated_coordinates =
[ [Longitude_Blue, Latitude_Blue], # Blue
point [Longitude_Orange, Latitude_Orange],
# Orange point [Longitude_Green,
Latitude_Green], # Green point [Longitude_Red,
Latitude_Red], # Red point [Longitude_Purple,
Latitude_Purple], # Purple point
[Longitude_Brown, Latitude_Brown], # Brown
point [Longitude_Pink, Latitude_Pink], # Pink
point [Longitude_Gray, Latitude_Gray], # Gray
point [Longitude_Olive, Latitude_Olive], #
Olive point [Longitude_Cyan, Latitude_Cyan], #
Cyan point ]
```

Occasionally, follow-up prompts were needed to ensure a position prediction was given for every coloured point.

2.9. Mapping – Localisation: real-world → map

In addition to the results reported in the main paper, we experimented with different grid sizes, grid line colours and map projections but were unable to find a configuration that improved the overall accuracy significantly. Other models showed an inability to correctly interpret the task and provide valid grid coordinates.

2.10. Flags

We examined GPT-4V's capacity for flag identification by presenting cropped segments from the original (from Sporcle) full grids of African and Asian flags. The crops correspond to the specific row and column layout indicated in the heatmaps (Fig. 6), which reflect the model's accuracy scores for each segment.

The models were provided with the following prompt with corresponding grid size parameters based on the crop size. The prompt directed the models as follows:

LLaVA-1.5 (*temperature*=0.2)

```
You will be playing the following game with
me:  Flags of {CONTINENT}.  Can you name
the flags of {CONTINENT}?  Respond in the
following format.  Think step by step during
this game.  Note that the grid is {ROWS} (row)
x {COLUMNS} (column).  Replace xxxx with
your guess of the flag.  RESPONSE (x row, y
column):  (1, 1) - xxxx, (1, 2) { xxxx
```

This focused approach aimed to discern GPT-4V's recognition efficacy when confronted with a condensed subset of visual information. Preliminary results indicate variability in the model's performance across different segments, with some showing high accuracy while others underperform. The heatmaps encapsulate this performance, highlighting areas of both strength and potential weakness within the model's identification capabilities.

While these findings might suggest regional biases in flag identification, conclusions about model bias should be drawn with caution. The variability in accuracy may also be influenced by factors such as image processing or the distinctive features of certain flags. As such, these results may warrant further investigation to better understand the underlying causes of performance discrepancies and to determine if they indeed reflect biases within the model's training data or recognition abilities.

References

[1] Rohan K. Geolocation - geoguessr images (50k), v1. https://www.kaggle.com/datasets/ubitquitin/geolocation-geoguessr-images-50k/, 2022. 3

[2] 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

[3] 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

[4] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation, 2022. 4

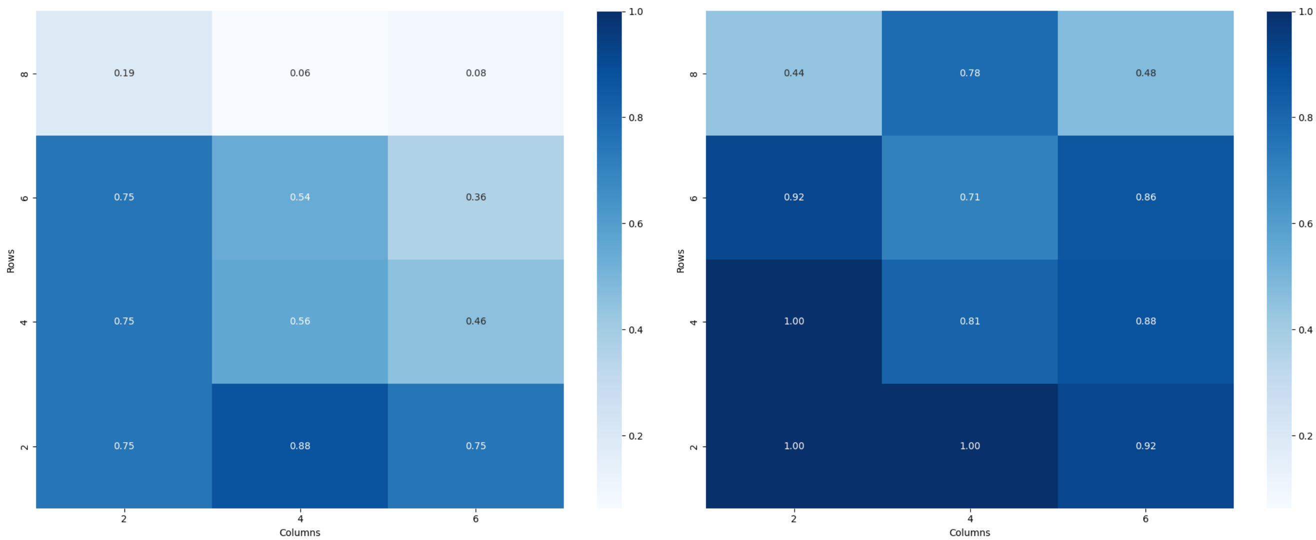


Figure 6. **Flag Identification Performance**. Heatmaps detail GPT-4V's accuracy in identifying flags for different sized cropped grid segments of Africa [left] and Asia [right]. Darker cells indicate higher identification accuracy.