NAACL ‘25

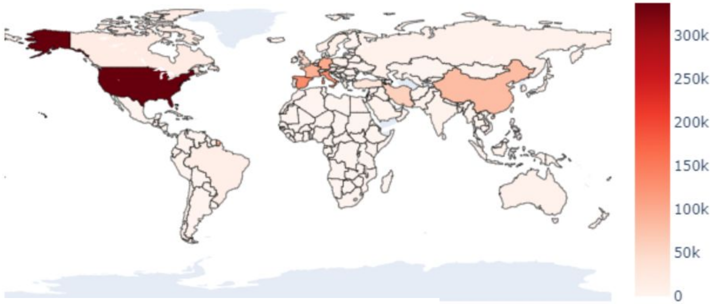
MapWise

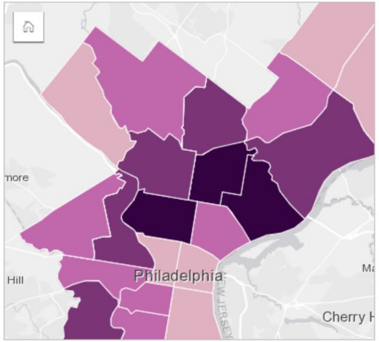
Evaluating Vision-Language Models for Advanced Map Queries

Srija Mukhopadhyay[1], Abhishek Rajgaria[2], Prerana Khatiwada[3],
Manish Shrivastava[1], Dan Roth[4], Vivek Gupta[5]

[1]International Institute of Information Technology Hyderabad, [2]University of Utah, [3]University of Delaware
[4]University of Pennsylvania [5]Arizona State University

Maps are Everywhere, But Can AI Read Them?





Maps are vital for representing complex geospatial data

VLMs handle photos well, but maps demand specific skills:

- Nuanced Spatial Reasoning
- Legend Interpretation
- Data Extraction & Correlation

Existing VQA benchmarks don't adequately test these map-specific challenges

The Gap in Map-Based VQA

General VQA: Doesn't focus on structured data visualizations

ChartQA: Focuses on charts, different visual grammar and knowledge requirements

MapQA (Chang et al., 2022): Good start, but limited scope (US-only, simpler templates, less map diversity)

What is the mustache made of?

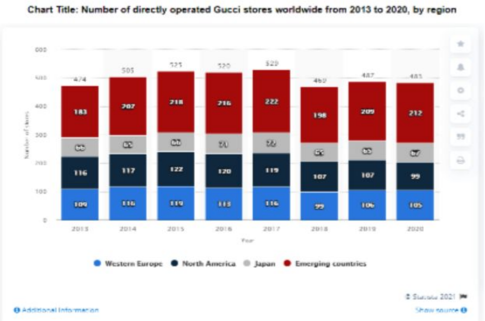
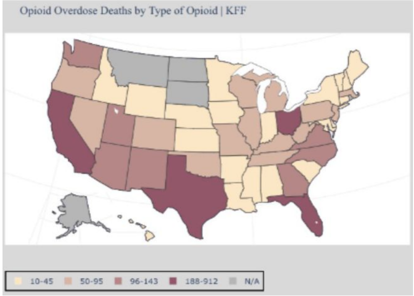


Chart Title: Number of directly operated Gucci stores worldwide from 2013 to 2020, by region

Which region has the least number of Gucci stores?



Name the regions with the value in the range 188-912?

We need a more comprehensive benchmark to drive progress

Introducing MapWise

Task: Advanced Question Answering on Choropleth Maps

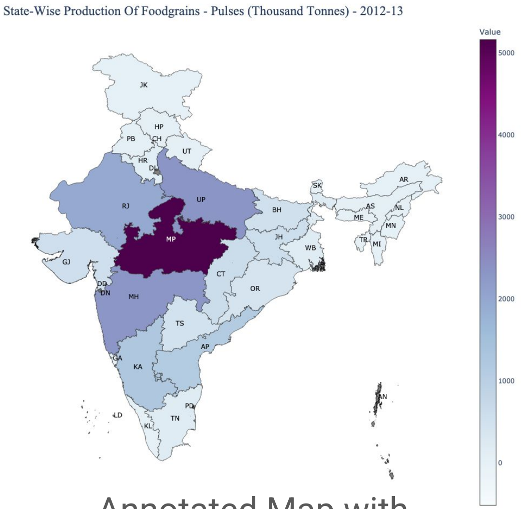
Regions: USA, India, China

Source Data: Real-world socio-economic/demographic stats

Question Complexity: 43 templates, all manually curated with an emphasis on:

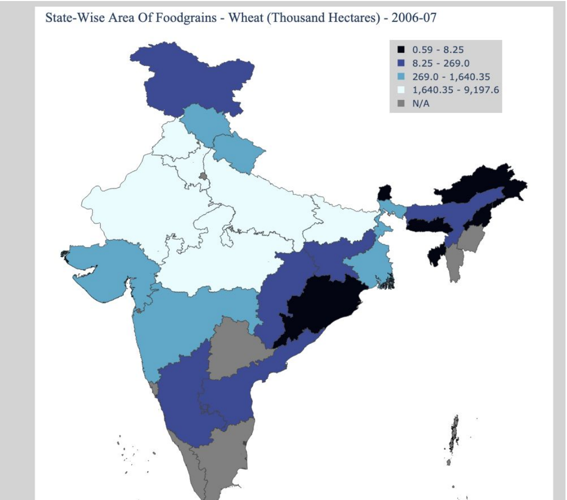
- Relative spatial reasoning ("bordering states", "easternmost")
- Data comparison & extraction ("higher value than X", "count states in range Y")
- Complex logical reasoning

A Rich and Diverse Dataset



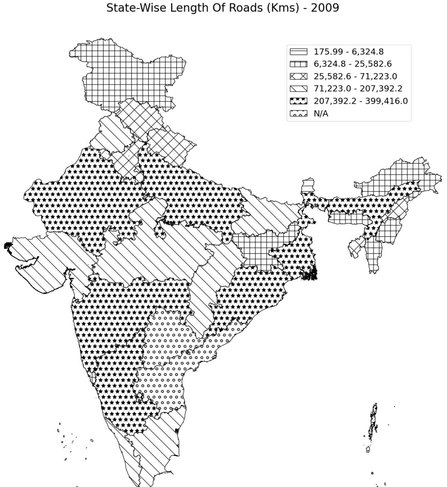
State-Wise Production Of Foodgrains - Pulses (Thousand Tonnes) - 2012-13

Annotated Map with Continuous Legend



State-Wise Area Of Foodgrains - Wheat (Thousand Hectares) - 2006-07

Unannotated Map with Discrete Legend



State-Wise Length Of Roads (Kms) - 2009

Hatched Map with Discrete Legend

We incorporate a diverse range of maps including annotated and unannotated maps, colored and hatched maps and maps with continuous and discrete legends

Incorporating Challenging Question Answering Tasks

Answer Type	Example Question
Binary	Yes or no: California is an outlier compared to its neighbours?
Single Word	Name the easternmost state that belongs to a higher value range compared to all its neighbors.
List	Which states in the East China Sea region have a value higher than the state Guangdong?
Range	What is the least value range in the west coast region?
Count	How many states bordering Canada have a value lower than New Mexico?
Ranking	Rank Rajasthan, Gujarat and Jammu & Kashmir in terms of legend values

Each question was created by an expert human, and then verified by another expert

Dataset Statistics

Our dataset includes around 1000 manually created questions for each country.

Map Type	USA	India	China
Continuous	33	51	49
Discrete	64	49	51

Answer Type	USA	India	China
Binary	449	456	441
Single Word	235	196	187
List	137	153	163
Range	130	103	112
Count	49	95	97
Ranking	30	29	26

Number of questions requiring the knowledge of Relative Regions

- USA - 145
- India - 206
- China - 214

Our Rigorous Evaluation Framework

Models Tested

- Closed-Source Multimodal Large Language Models

  GPT-4o, Gemini 1.5 Flash

- Open-Source Vision Language Models:

  CogAgent-VQA, InternLM-XComposer2, Idefics2, QwenVL

All models were compared against a Human Baseline

- Dataset baselined by 3 expert humans
- 450 representative questions to test validity

Robust evaluation through Prompting Strategies

Prompting Strategies

- Zero-Shot Chain-of-Thought (CoT)

  Urging the model to think step by step and reason before giving the final answer

- Explicit Extraction & Reasoning (EER)

  Step 1: Extraction of Regions

  Step 2: Extraction of Relevant Places

  Step 3: Extraction of Values from Legend

  Step 4: Reasoning based on Extracted Values

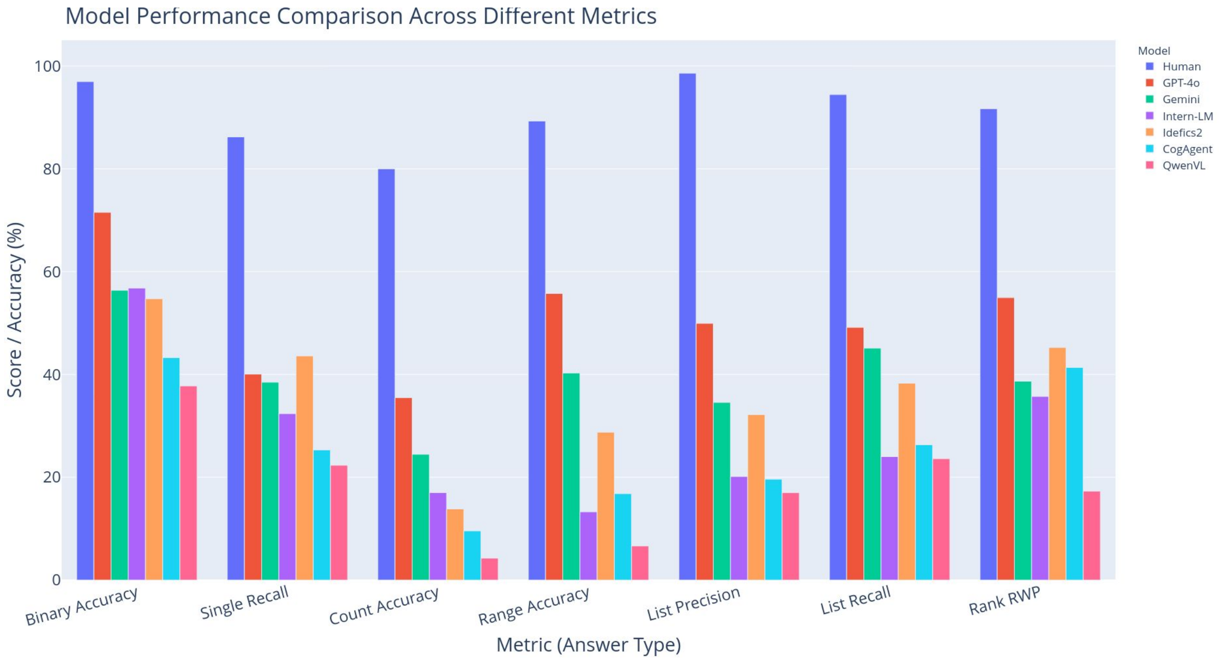
Tailored Evaluation Metrics

Evaluation Metric

Answer Type	Evaluation Metric
Binary	Exact Match and Accuracy Scores
Single Word	Recall as some questions had multiple correct answers
List	Precision and Recall
Range	Relaxed Accuracy to account for differences in continuous maps
Count	Exact Match and Accuracy Scores
Ranking	Average of precision for each rank (Rank Wise Precision)

Our evaluation metric is tailored for each answer type to ensure fairness

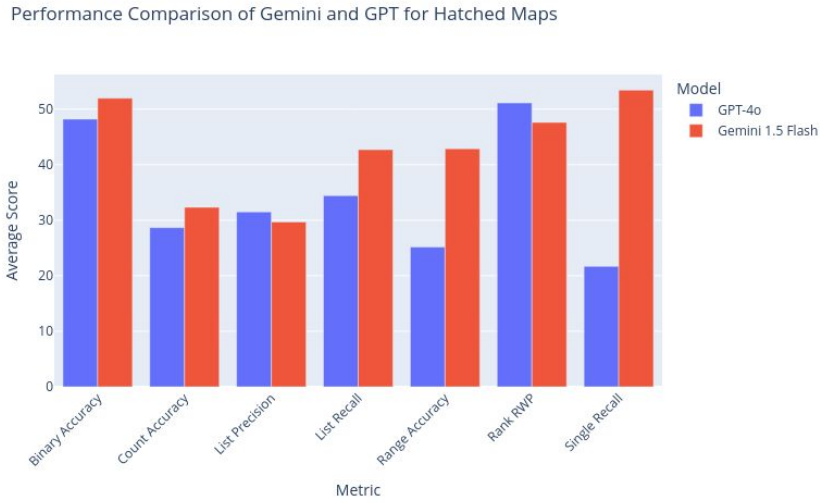
Significant Gap Between State-of-the-Art Models and Human Performance

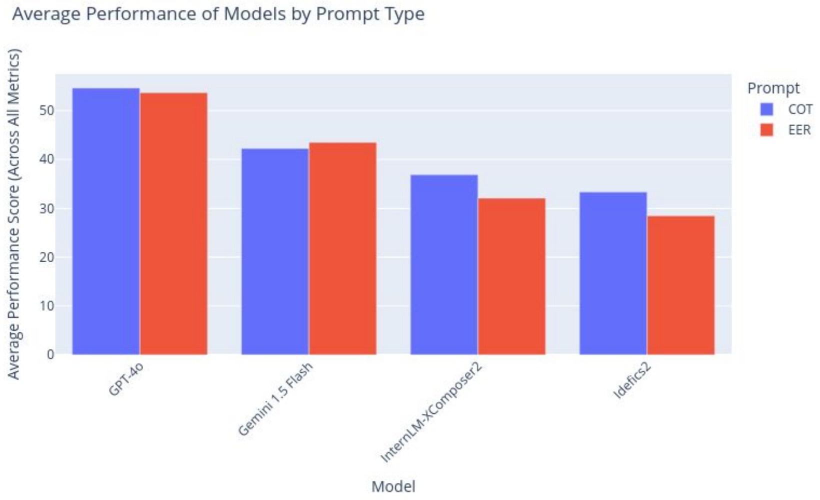


Even top models struggle significantly (~50% gap on List/Count tasks)

Closed-Source Leads, Prompting Strategy Matters

- GPT-4o/Gemini >> Open-Source Models
- Gemini excels on hatched maps
- CoT leads to generally better results, but Gemini follows EER instructions well.
- Larger models sometimes inherently adopted a step-wise approach


Performance Comparison of Gemini and GPT for Hatched Maps


Average Performance of Models by Prompt Type

Performance Varies Across Map and Question Types

What was difficult for models?

Easier	Harder
Binary, Single Word Questions	Count, List, Ranking Questions
Discrete, Colored Maps	Continuous, Hatched Maps

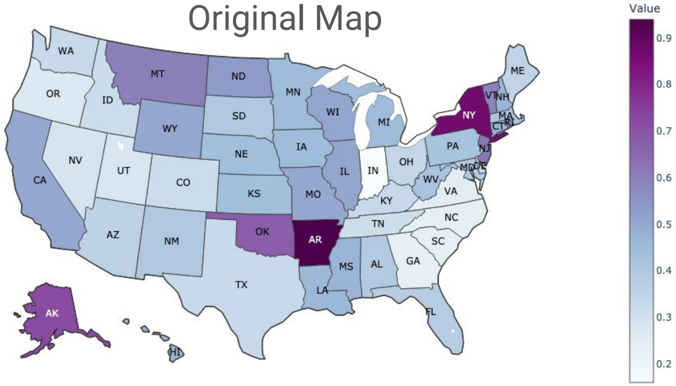
Did annotations help?

To an extent, but minimally

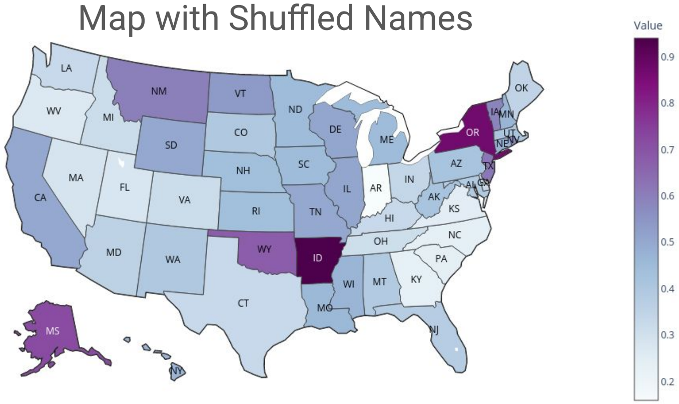
Did regions affect performance?

Some variations observed (potential training data bias?)

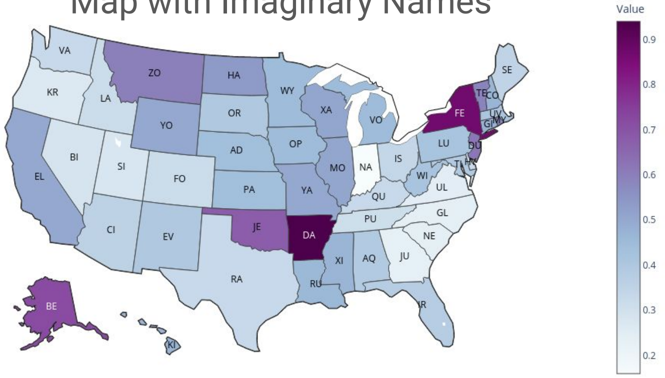
Do Models REALLY See the Map?



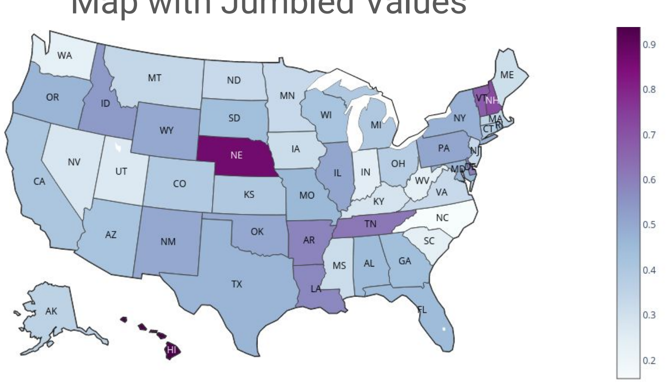
Original Map



Map with Shuffled Names



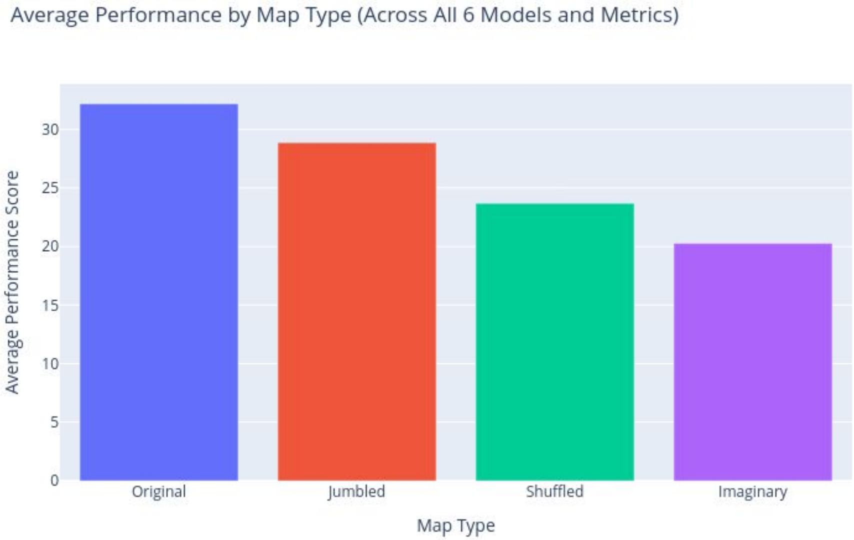
Map with Imaginary Names



Map with Jumbled Values

Counterfactual Analysis Says... Sometimes

- Performance drops significantly with imaginary/shuffled names
- Open-source models struggle drastically here
- Models handle jumbled values slightly better but still show decline.



Closing Notes

Recap

Introduced MAPWise, a diverse and challenging benchmark for advanced map querying.

Key Finding

Revealed significant limitations in current VLM reasoning and grounding abilities on choropleth maps.

Call to Action

MAPWise provides a valuable resource to measure and drive progress in making AI truly geospatially aware.

Thank you!

To know more: https://map-wise.github.io/