MAPWise: Evaluating Vision-Language Models for Advanced Map Queries

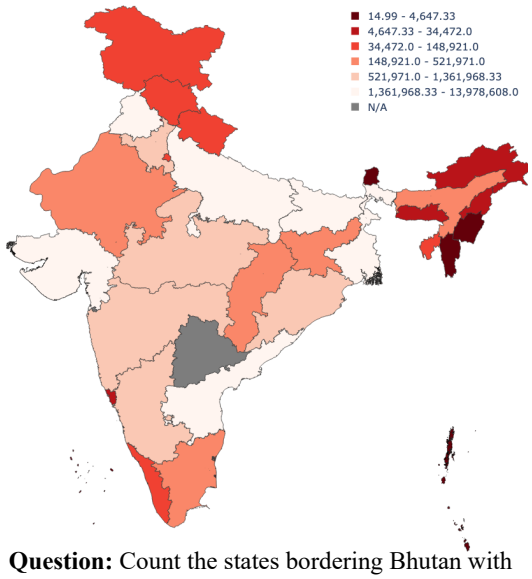
Srija Mukhopadhyay[1], Abhishek Rajgaria[2], Prerana Khatiwada[3]
Manish Shrivastava[1], Dan Roth[4], Vivek Gupta[5]*
[1]IIIT Hyderabad, [2]University of Utah, [3]University of Delaware
[4]University of Pennsylvania, [5]Arizona State University
srija.mukhopadhyay@research.iiit.ac.in, abhishek.rajgaria@utah.edu, preranak@udel.edu
m.shrivastava@iiit.ac.in, danroth@seas.upenn.edu, vgupt140@asu.edu

Abstract

Vision-language models (VLMs) excel at tasks requiring joint understanding of visual and linguistic information. A particularly promising yet under-explored application for these models lies in answering questions based on various kinds of maps. This study investigates the efficacy of VLMs in answering questions based on choropleth maps, which are widely used for data analysis and representation. To facilitate and encourage research in this area, we introduce a novel map-based question-answering benchmark, consisting of maps from three geographical regions (United States, India, China), each containing around 1000 questions. Our benchmark incorporates 43 diverse question templates, requiring nuanced understanding of relative spatial relationships, intricate map features, and complex reasoning. It also includes maps with discrete and continuous values, covering variations in color mapping, category ordering, and stylistic patterns, enabling a comprehensive analysis. We evaluated the performance of multiple VLMs on this benchmark, highlighting gaps in their abilities, and providing insights for improving such models. Our dataset, along with all necessary code scripts, is available at map-wise.github.io.

1 Introduction

Vision-Language Models (VLMs) have demonstrated impressive capabilities in tasks requiring joint understanding of visual information and natural language. They have achieved significant success in areas like image generation (Bie et al., 2023), multimodal sentiment analysis (Lai et al., 2023) and visual question answering (VQA) (Antol et al., 2015; Krishna et al., 2017; Kabir et al., 2024; Masry et al., 2022). However, when applied to map-based question answering, the reasoning abilities of these models remain largely unexplored (Chang et al., 2022).



Question: Count the states bordering Bhutan with values in range below 34472.0?
Answer 2

Figure 1: A question-map pair from our MAPWise dataset and the corresponding gold truth answer.

Choropleth maps, which use varying shades or colors to represent geographical data, present a unique challenge (Chang et al., 2022). While humans can readily grasp the spatial patterns and information conveyed by these color variations, their interpretation poses a significant challenge for visual language models and other analytical tools. This difficulty arises from the inherent challenge of translating visual data represented by different colors or shades into simpler, tabular formats.

This research addresses this gap by analyzing the performance of VLMs in answering questions related to choropleth maps representing different geographical regions (Example shown in Figure 1). We aim to answer the following research questions:

($RQ1$) How effectively can VLMs answer questions about Choropleth maps of different geographical regions?

($RQ2$) What prompting strategies can improve

*corresponding author.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 9348–9378
April 29 - May 4, 2025

the performance of models for Map Visual Question Answering (Map-VQA)?

($RQ3$) What biases are present in these models with regards to Map VQA?

($RQ4$) How effectively do these models attend to the provided map when performing visual question-answering tasks?

To address these research questions, we created a novel dataset, **MAPWise**, specifically for map-based VQA. This dataset comprises **1,000** questions for three geographical regions: the United States, India, and China. The questions were manually created based on **43** unique templates, designed to evaluate model capabilities across topics ranging from data extraction to complex reasoning.

Furthermore, the dataset includes various map representations, including maps with and without annotations, a diverse range of colormaps, and stylistic patterns like hatching, creating a robust benchmark. We have used this dataset for experimentation across various leading VLMs and MMLMs, using diverse prompting techniques to establish a viable baseline. Our study also included an analysis of model performance on counterfactual maps. These maps featured imaginary state names, jumbled state names, and counterfactual statistics. Our analysis aimed to not only understand how well the models relied on the provided map data but also to what extent they relied on their internal knowledge. Our contributions are:

- ***Dataset*:** The MAPWise dataset, tailored for choropleth maps, provides diverse questions that test various aspects of geographical and spatial understanding.
- ***Models*:** Baseline performances using VLMs provide a reference point for research in map-based VQA. We also included human baseline scores for a more comprehensive analysis.
- ***Bias and Counterfactual Analysis:*** In-depth analysis of biases present in the models and our counterfactual analysis highlights areas of struggle and offers insights for improvement.

2 The MAPWise Dataset

This section details the creation process of the MAPWise dataset, including data gathering, manual question creation, and dataset validation.

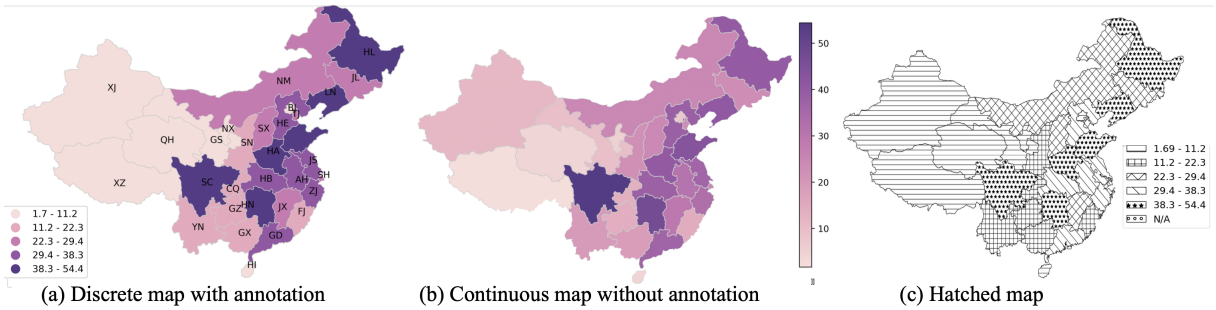
2.1 Dataset Creation

Data Sources. The **MAPWise** dataset was created using data from three countries: India, USA, and China. We have meticulously chosen reliable sources to gather socioeconomic and demographic statistics for each country, as described below.

- For **India**, we sourced data from the Reserve Bank of India's *"Handbook of Statistics on Indian States."* This resource provides extensive data across various periods, including details such as state-wise cold storage capacity, rural population figures, and the area of non-food grains like cotton.
- For **USA**, the primary data source was the *"Kaiser Family Foundation"*, which specializes in healthcare statistics. This includes information on health insurance coverage for adults without dependent children, age-adjusted suicide rates, and weekly COVID-19 vaccine allocations.
- For **China**, we obtained data from the *"National Bureau of Statistics of China"*, which provides data on household consumption expenditure, urban unemployment rates and natural growth rate.

Map Variations. The dataset consists of maps representing data in two primary forms: discrete, where the legend is divided into distinct groups and continuous, where the legend is distributed over a spectrum. The maps also include variations in the presence or absence of annotations, which provide additional contextual information. Our dataset also has maps with black-and-white textured patterns or hatches for discrete data, different colormap variations (light, dark, and gradient scales), and varying paper background colors (white and grey). These variations test the models' capability to handle diverse visual presentations. We generated maps with annotations, without annotations, and with hatching for each country using the Plotly library. Examples of the generated maps can be found in Figure 2.

Question Generation. To create a comprehensive and insightful benchmark, we designed question templates with varying levels of difficulty, ranging from simple *yes/no* questions to more complex *region association* questions that required reasoning based on relative locations.



(a) Discrete map with annotation (b) Continuous map without annotation (c) Hatched map

Figure 2: Examples of map with annotations, without annotations for the same underlying data. Additionally, hatched maps were created to better assess model understanding and performance.

The dataset includes three major question types: Binary questions, which require a simple yes or no answer based on the map; Direct Value Extraction questions, which ask for a specific numerical or nominal value related to a particular region or the legend; and Region Association questions, which involve identifying or counting regions meeting some specific criteria, often requiring geospatial reasoning and reasoning about relative regions.

Answer Type	Example Question
Binary	Yes or no: California is an outlier compared to its neighbours?
Single Word	Name the eastern-most state that belongs to a higher value range compared to all its neighbours.
List	Which states in the East China Sea region have a value higher than 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 the legend value in region bordering Pakistan.

Table 1: Example questions along with the different types of possible answers.

Each question could have answers in one of the following formats: Binary (Yes/No), Single Word, Count, List, Range, and Ranking. Examples of these are shown in Table 1. All questions were manually created by expert annotators, with the help of provided templates, with 10 questions created for each map. Overall, we created around 1000 question-answer pairs for each country. The final dataset statistics are presented in Table 2.

Dataset Validation. The generated questions were initially validated by expert annotators (detailed in Appendix B). Following that, we carried out a process of human evaluation that played a critical role in confirming the accuracy of our dataset. It also served as a benchmark for comparing model performance. Table 9 presents human evaluation metrics for the three countries.

Type	Country	USA	India	China
Maps	Total	97	100	100
	Continuous	33	51	49
	Discrete	64	49	51
Answer Types	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
Question	Relative regions	145	206	214

Table 2: Overview of MAPWise statistics.

3 Experimental Evaluation

This section outlines our experimental setup: we selected a mix of closed-source and open-source Vision-Language Models (VLMs) and Multimodal Large Language Models (MLLMs) for a comprehensive analysis. These models were tested with various prompting techniques, and we developed an evaluation metric to assess different answer types.

3.1 Baseline Models

Closed-Source MLLMs. For analysis on closed source models, we used Gemini 1.5 Flash (Gemini, 2024) and GPT-4o (OpenAI, 2024). These models are known for their advanced features and proprietary implementations.

Open-Source VLMs. We selected CogAgent, InternLM-XComposer2, Idefics 2, and Qwen VL. CogAgent-VQA (Hong et al., 2024) is an 18-billion-parameter VLM specializing in GUI understanding and navigation. InternLM-XComposer2 (Dong et al., 2024), an adaptation of InternLM2-7B (InternLM2, 2024), excels in producing high-quality long-text multimodal content and reasoning within visual-language understanding contexts.

QwenVL (Bai et al., 2024), a generalist 7-billion-parameter VLM built on top of Qwen-LM (Qwen, 2023), uses adapted visual encoders and general and multi-task pretraining. These models were chosen due to their accessibility and contributions to the research community, each offering distinct approaches to processing and interpreting visual information.

3.2 Prompting Strategies

We evaluated the baseline models under two distinct prompting settings:

1. ***Zero-Shot Chain-of-Thought Prompting (COT)***. We leverage the Chain-of-Thought (Wei et al., 2024) prompting, encouraging it to reason through the steps leading to the answer, when provided with a map and a question.

2. ***Explicit Extraction and Reasoning (EER)***. Here, we created a custom prompt that explicitly outlined the reasoning steps the model should follow to answer the specific question. This prompt was broken down into four distinct reasoning steps:

   - *Extraction of Regions*. The model was prompted to identify the regions whose data was required to answer the question.

   - *Extraction of Relevant Places*. The model was then instructed to extract the specific locations associated with the identified regions.

   - *Extraction of Values from Legend*. The model was then directed to extract the values corresponding to the extracted regions from the map's legend.

   - *Reasoning based on Extracted Values*. Finally, the model was prompted to reason based on the extracted values to reach the answer.

   This approach helped break down the reasoning process into smaller, more manageable steps, preventing the model from becoming overwhelmed and guiding it towards a more focused and structured reasoning process.

During the evaluation, all models were given the same prompt in order to fairly and consistently assess their ability to reason. The prompts used have been presented in the Appendix.

3.3 Evaluation Details

The evaluation process adapts to various answer types within in the dataset by employing tailored metrics and criteria for each specific answer type. Additionally, normalization was applied wherever necessary to ensure consistency and accuracy in the assessment.

For binary *yes/no* and integer *count* answers, we implemented an exact match criterion and accuracy as the evaluation metric. For single-word answers, as some questions have multiple applicable responses, we employed the recall metric for better evaluation. For state names, a valid answer could be either a two-digit state code or the full state name. For ranges, we first normalized the ranges to absolute values (e.g. *1k to 1000*) and then compared them. For discrete maps, only exact match was expected, whereas for continuous maps, we gave a full score of 1 for exact match and a partial score of 0.5 for overlapping responses.

For list type answer, we used precision and recall metrics because predicted lists often contained irrelevant states (*false positives*) and missed relevant states (*false negatives*).

For rank-type answers, we prompted the model to assign ranks to states based on map values. However, due to the difficulty in accurately distinguishing shades, models frequently assigned states to wrong shades, resulting in multiple states sharing the same rank despite differing shades. Additionally, for some questions, ground truth involved multiple states in the same rank because of states having identical shades or patterns. To evaluate this, we designed a "Rank-wise Precision (RWP)" method, computing precision for each rank and then averaging across all ranks. We also evaluated other ranking metrics, including Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP), as detailed in Appendix C.

Note for Open Source VLMs. Smaller models, like QwenVL, CogAgent-VQA, and InternLM-XComposer2, faced challenges in producing answers in the desired format. To address this, we used an **"LLM as an Extractor"** approach, using Gemini 1.5 Flash to extract answers from their outputs. Manual verification of *150* samples confirmed that Gemini 1.5 Flash primarily acted as a extracting and formatting tool, preserving the original model's answer in *138* cases. In the remaining *12* cases, the original model had not clearly answered the question, for which Gemini 1.5 Flash reported *"Answer cannot be extracted"*.

4 Results and Analysis

MAPWise: A Challenging Benchmark. The MAPWise dataset presents a compelling benchmark for evaluating the reasoning abilities of current Vision-Language Models (VLMs). As shown in Table 3, models consistently perform significantly worse than the human baseline, particularly with questions requiring intricate reasoning, such as counting or providing a list of regions where the difference in scores is close to **50%** on average. This substantial performance gap highlights a significant limitation in the reasoning capabilities of existing VLMs, underscoring the need for further research to bridge this gap. All the results we obtained are presented in the Appendix H.

Model	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall	Rank RWP
Human	96.97	86.21	80.00	89.29	98.61	94.44	91.67
GPT-4o	**71.52**	40.06	**35.48**	**55.75**	**49.94**	**49.17**	**54.94**
Gemini	56.36	38.49	24.47	40.27	34.55	45.11	38.69
Intern-LM	56.80	32.37	17.02	13.27	20.14	24.02	35.71
Idefics2	54.71	**43.59**	13.83	28.76	32.19	38.29	45.24
CogAgent	43.27	25.32	9.57	16.81	19.62	26.32	41.36
QwenVL	37.75	22.33	4.26	6.64	17.00	23.60	17.31

Table 3: Results for different models when evaluated on annotated maps of India using the zero-shot COT prompt, compared against the human baseline.

Model Performance Comparison. While model performance varied across different answer types and countries, GPT-4o consistently emerged as the top performer in most categories, closely followed by Gemini 1.5 Flash (as shown in Table 3). Notably, Gemini 1.5 Flash demonstrated superior performance on hatched maps (as seen in Table 4), likely due to its stronger legend resolution and data extraction capabilities. However, GPT-4o's robust reasoning skills generally led to better scores across other task types.

Model	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall	Rank RWP
			USA				
Gemini 1.5 Flash	49.36	56.20	51.22	53.95	20.51	35.35	43.52
GPT-4o	49.78	16.14	26.83	26.67	26.96	32.60	60.19
			India				
Gemini 1.5 Flash	52.75	48.65	23.53	34.38	38.95	47.33	38.89
GPT-4o	49.72	28.38	31.37	30.77	34.19	36.27	53.57
			China				
Gemini 1.5 Flash	53.80	55.41	22.22	40.32	29.61	45.41	60.42
GPT-4o	45.11	20.56	27.78	18.03	33.33	34.40	39.58

Table 4: Scores for Gemini 1.5 Flash and GPT-4o for hatched maps using zero shot COT prompt.

While open-source models generally lag behind their closed-source counterparts in performance, Idefics2 and InternLM-XComposer2 demonstrate surprisingly strong results. However, we observed that open-source models struggle significantly with questions requiring complex reasoning, with QwenVL achieving a low 4.26% accuracy on tasks involving counting. This stark difference underscores the crucial need for models not only to excel in data extraction but also to possess sophisticated reasoning skills, particularly in the domain of geo-spatial reasoning.

Prompt Effectiveness. While most models consistently perform better with the standard Chain-of-Thought (COT) prompt compared to the Explicit Extraction and Reasoning (EER) prompt (as evident in Table 5), a notable exception is Gemini 1.5 Flash, which performs comparably to the EER prompt. This suggests that Gemini 1.5 Flash possesses particularly strong instruction-following capabilities. Smaller, open-source models likely struggle with following the complex, step-wise instructions within the EER prompt. However, analysis of responses from larger models reveals that they implicitly adopt a methodology similar to EER, demonstrating impressive progress in their reasoning abilities and mimicking human-like thinking.

Prompt	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall	Rank RWP
			GPT-4o				
COT	66.97	47.53	50.52	59.40	53.93	57.56	46.58
EER	63.33	60.65	45.36	59.83	43.47	46.48	56.62
			Gemini 1.5 Flash				
COT	62.27	51.83	13.40	52.97	22.76	38.96	53.63
EER	61.50	54.09	24.74	52.14	23.01	39.54	49.54
			InternLM-XComposer2				
COT	54.09	50.54	21.65	34.32	21.67	29.91	46.15
EER	53.86	29.25	18.56	28.39	26.63	39.78	28.21
			Idefics2				
COT	54.09	38.39	19.59	28.81	22.98	28.02	41.67
EER	42.50	23.66	21.65	24.15	20.97	24.74	41.67

Table 5: Performance of different models across prompting strategies. The models were evaluated on annotated maps of China.

5 Biases in Model Prediction

This section analyzes the performance variations of models across different map and question variants. These observations are influenced by question type, but we highlight the most prominent insights.

5.1 Map Variants

Discrete vs. Continuous Maps. While it is challenging to directly compare model performance on continuous and discrete maps due to the differing question types, a general trend emerges: models tend to perform better on discrete maps (as shown in Table 6). This trend is particularly pronounced for questions involving counting and extracting

ranges, suggesting that models might struggle with accurately extracting legend ranges and color resolution in continuous maps. Interestingly, models performed significantly better on single-word answers within the continuous category. This may be attributed to the simplicity of these questions, as the task itself is inherently challenging for humans.

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall	Rank RWP
GPT-4o							
continuous	64.05	62.12	25.00	61.84	36.70	42.11	56.94
discrete	73.72	35.62	56.10	74.76	39.85	47.80	43.98
hatched	49.78	16.14	26.83	26.67	26.96	32.60	60.19
Gemini 1.5 Flash							
continuous	63.03	56.52	25.00	43.75	38.84	53.39	43.98
discrete	66.20	53.64	56.10	70.48	38.66	50.26	60.34
hatched	49.36	56.20	51.22	53.95	20.51	35.35	43.52
InternLM-XComposer2							
continuous	54.55	50.72	12.50	22.50	23.56	34.18	33.33
discrete	51.76	27.91	36.59	19.05	22.08	32.41	48.15
hatched	45.49	31.40	53.66	17.11	25.84	31.31	36.11
Idefics2							
continuous	61.21	63.77	12.50	27.50	23.55	41.24	30.56
discrete	51.06	59.69	19.51	30.48	28.48	44.02	50.00
hatched	52.79	56.98	12.20	25.00	22.99	42.09	45.37

Table 6: Performance of different models across annotated discrete, annotated continuous and hatched maps of USA, using the zero-shot COT prompt.

Colored Maps vs. Hatched Maps. All models consistently performed better on colored maps compared to hatched maps, demonstrating a preference for colored depictions of data (as seen in Table 6). This trend is notable, as even models like GPT-4o experienced significant score drops on hatched maps, highlighting a lack of robustness. Impressively, Idefics2 displayed the least performance decline, suggesting a more robust ability to accurately extract data from these visually complex maps.

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall	Rank RWP
GPT-4o							
with	71.52	40.06	35.48	55.75	49.94	49.17	53.70
without	66.45	40.92	30.85	53.54	46.23	47.09	55.56
Gemini 1.5 Flash							
with	56.36	38.49	24.47	40.27	34.55	45.11	38.69
without	58.99	37.39	23.40	35.84	36.25	46.21	45.83
InternLM-XComposer2							
with	56.80	32.37	17.02	13.27	20.14	24.02	35.71
without	53.51	34.08	14.89	13.27	27.84	35.84	39.29
Idefics2							
with	54.17	43.59	13.83	28.76	32.19	38.29	45.24
without	56.58	43.48	15.96	23.45	28.04	36.26	48.81

Table 7: Performance of different models across maps of India with and without annotations, using the zero-shot COT prompt. Here, "with" and "without" represent the presence and absence of annotations respectively.

Maps with and without annotations. As shown in Table 7, models generally exhibited similar performance on maps with and without annotations, with only a slight improvement observed for annotated maps in some cases. Surprisingly, we also found instances where models performed better on maps without annotations. This suggests that while annotations can be beneficial, they are not a critical factor in building models for understanding maps.

5.2 Country-Wise Performance

Table 8 presents model performance across different countries. While a consistent pattern is difficult to discern, a notable trend emerges: open-source models generally demonstrate consistent performance across countries, while closed-source models exhibit greater variation. The exact cause of this variation remains unclear, but potential contributing factors include biases in the training data.

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall	Rank RWP
GPT-4o							
USA	70.26	44.68	51.02	71.28	38.64	45.61	49.17
India	71.52	40.06	35.48	55.75	49.94	49.17	54.94
China	66.97	47.53	50.52	59.40	53.93	57.56	45.66
Gemini 1.5 Flash							
USA	65.03	54.65	51.02	63.10	38.73	51.43	53.80
India	56.36	38.49	24.47	40.27	34.55	45.11	38.99
China	62.27	51.83	13.40	52.97	22.76	38.96	54.31
InternLM-XComposer2							
USA	52.78	35.86	32.65	20.00	22.63	33.07	42.22
India	56.80	32.37	17.02	13.27	20.14	24.02	35.71
China	54.09	38.39	19.59	28.81	22.98	28.02	41.99
Idefics2							
USA	54.79	61.11	18.37	29.66	26.64	42.99	42.22
India	54.17	43.59	13.83	28.76	32.19	38.29	45.24
China	54.09	50.54	21.65	34.32	21.67	29.91	46.15

Table 8: Performance across annotated maps of USA, India, China, using the zero-shot COT prompt.

5.3 Analysis across Question and Answer Types

Table 8 shows that models performed best on questions requiring a binary answer, followed by single-word answers, highlighting their strong data extraction capabilities. Closed-source models like Gemini 1.5 Flash and GPT-4o also excelled at questions expecting a range; however, smaller models struggled in this domain, likely due to limited reasoning or color extraction skills. Models encountered the most difficulty with tasks requiring a count or listing, which demand complex reasoning, external knowledge, and geospatial understanding. These questions proved challenging not only for models but also for humans (as shown in Table 9). For questions concerning relative regions, models struggled with most categories apart from binary, further highlighting the complexity of these tasks, which require external knowledge, relative region extraction, and complex reasoning. Models especially struggled with List or Range based questions, with smaller models suffering the most (as seen in the Appendix H tables for relative regions).

5.4 Comparison with Random Baselines

Furthermore, we observed a surprising trend where some models performed below the random baseline. The follow section highlights the possible reasons behind this anomaly.

Binary Questions: CogAgent and QwenVL demonstrated performance significantly below a random baseline when answering binary-type questions. This reduced accuracy primarily stems from a higher incidence of irrelevant responses and repeated token generation by these models. Specifically CogAgent failed to provide a processable answer for 78 binary-type questions, and these responses were insufficient for our LLM-based extraction method. QwenVL failed in 97 instances.

These instances represent more than simply under-performance; they constitute a failure to generate any valid response at all. In our evaluation, these cases were treated as negative outcomes (i.e., the models did not produce the expected answers), resulting in notably low accuracy scores.

List Precision: Similar issues presented a critical challenge for list precision. The models often failed to provide relevant state names, or sometimes, provided none at all. Furthermore, the observed responses indicated poor precision in data processing and a limited grasp of the underlying tasks.

Example 1: Value Deduction. Consider the question: "Which state in Central India has the furthest value to Andhra Pradesh?" The ground truth answer is "Chhattisgarh". CogAgent's response was "Madhya Pradesh". The actual observation of the data revealed that Andhra Pradesh has a value ">4000", Madhya Pradesh has "1000-1500", and Chhattisgarh has "<500".

This response demonstrates the model's failure to accurately deduce values from the given map data and therefore, to identify the correct answer.

Example 2: Task Comprehension and Hallucination. Another example highlights the model's inability to comprehend the task requirements. The question was: "Which states that share an international land border have a value similar to Madhya Pradesh?" The ground truth answer includes: "Rajasthan, Punjab, Sikkim, Arunachal Pradesh, Nagaland, Manipur, Mizoram, Tripura, Meghalaya". However, CogAgent's response was: "Nepal, Bangladesh, Pakistan, Bhutan, China". This clearly reveals the model's failure to comprehend the task, as it incorrectly identifies countries rather than the required states.

The response also showcases how the model might hallucinate and generate answers, reducing its overall scores. These instances are consistent with the observed challenges these models had on the binary questions. These examples collectively demonstrate the model's limitations in both understanding the task requirements and accurately extracting and correlating data.

6 Human Evaluation and Baseline

We conducted a human evaluation of the MapWise dataset to establish a human baseline and benchmark model performance against human evaluators. The dataset presents significant challenges, requiring evaluators to identify subtle shades and patterns while demonstrating understanding of spatial geographical relationships. Our evaluation encompassed 450 questions, comprising 150 unique questions uniformly sampled from each of three countries. For each country, the sample included approximately 75 maps and 40 templates. To ensure comprehensive coverage, we maintained an equal distribution across answer types and map categories, incorporating both continuous and discrete maps as well as relative region-type questions. This systematic approach was implemented consistently across all three countries to capture the full range of scenarios within the dataset. For validation, we employed a majority voting system among three independent annotators.

Country	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall	Rank RWP
USA	94.74	96.67	88.89	100.00	95.16	93.55	92.59
India	96.97	86.21	80.00	89.29	98.61	94.44	91.67
China	100.00	88.99	79.31	80.77	79.76	79.76	80.00

Table 9: Human Baseline results (in %).

As shown in Table 9, the less-than-perfect human performance highlights the complexity of the task and offers a realistic benchmark against which model performance can be compared. Several common challenges contribute to the dataset's complexity, even for human evaluators. These include confusing color shades, particularly in continuous maps, numerous range groups in discrete maps, difficulty in understanding patterns for hatched maps and the challenge of accurately interpreting values for regions with smaller areas.

From Table 10, we observe that for binary, range and list type answer, there is nearly 100% majority agreement among human evaluators. However,

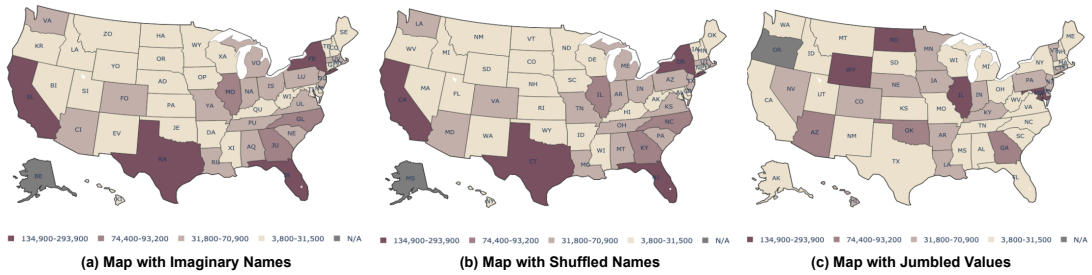


Figure 3: Examples of map with Imaginary and Shuffled names and Jumbled Values for the same underlying data.

Country	Binary (yes/no)	Single	Count Integer	Range A-B, >A, <B	List	Rank
USA	100.00	96.67	88.89	100.00	96.77	100.00
India	96.97	89.66	86.67	100.00	100.00	100.00
China	100.00	96.43	89.66	100.00	100.00	80.00

Table 10: Percentage of responses which aligns with the Majority voted response

there is a slight decline in majority agreement for single type answers and least majority for count type answer, highlighting the confusion and variability in responses among human evaluators.

7 Experiments with Counterfactual data

We performed additional analysis to evaluate models which are trained extensively on large datasets, under conditions where their internal factual knowledge was limited. To carry out our analysis, we created three types of counterfactual data that forced the models to rely exclusively on the provided maps. Figure 3 shows such counterfactual maps.

For the counterfactual dataset generation, we first uniformly sampled a subset of 240 unique Questions from USA dataset, spreading over 90 Maps and 26 Templates. We also ensured approximately equal distribution of each answer type. Using the sampled dataset as a representative sample (consisting of original names and values), we applied the following modifications to create our counterfactual dataset:

Imaginary names. States were assigned imaginary names, generated by GPT-4. (*e.g., Alabama was renamed Aquilis, Arkansas became Davina, etc.*) The first two letters of the imaginary names were used as state codes for the annotated maps.

Shuffled names. The names of different US states were randomly shuffled while retaining the values of each geographical region. Annotated maps with these shuffled state codes were generated (e.g. Alabama became Montana, Arkansas became Idaho).

Jumbled values. The values corresponding to each of the different US states were shuffled, keeping the legend fixed. As a result, several question answer pairs needed to be re-evaluated.

CF Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall
Gemini 1.5 Flash						
Original	59.18	35.42	20.00	35.42	47.52	60.20
Imaginary	53.06	23.96	11.11	35.42	24.86	38.27
Shuffled	63.27	25.00	22.22	37.50	18.84	25.68
Jumbled	53.06	30.21	31.11	40.63	39.88	45.41
GPT-4o						
Original	61.12	37.48	22.11	36.99	49.58	62.25
Imaginary	55.09	25.98	13.13	37.46	26.89	40.29
Shuffled	65.31	27.03	24.24	39.53	20.87	28.72
Jumbled	55.09	32.24	33.13	42.67	41.92	47.45
Idefics2						
Original	55.10	31.25	13.33	30.21	26.19	47.79
Imaginary	46.94	0.00	8.89	16.67	0.00	0.00
Shuffled	53.06	12.50	13.33	25.00	7.82	13.95
Jumbled	32.65	14.58	11.11	23.96	25.83	43.88
InternLM-XComposer2						
Original	46.94	13.54	28.89	14.58	26.17	40.82
Imaginary	53.06	0.00	20.00	8.33	3.96	9.86
Shuffled	55.10	10.42	15.56	13.54	11.85	15.31
Jumbled	42.86	13.54	15.56	6.25	22.59	30.78

Table 11: Counter Factual Results (in %) for zero-shot COT prompt. CF represents Counter Factual

Adjustments to the prompts were made in accordance with the specific requirements of each counterfactual dataset. For example when dealing with imaginary names, the following instruction was included: *"The map in the image represents fictional names for each state as specified in the following dictionary. Use this dictionary while analyzing the map"*. A corresponding dictionary was provided for reference within the prompt. Table 11 presents the results for Gemini 1.5 Flash, GPT-4o, Idefics2 and InternLM-XComposer2, evaluated using the zero-shot COT prompt (Appendix A for contains results for the remaining models and the EER prompt). At a high level, it is evident that the closed source model consistently outperformed the open-source models across all three types of counterfactual datasets.

Upon closer inspection, we notice a significant decline in performance for Single and List type answers when using imaginary and shuffled names compared to the original dataset. However, the comparable or better results for Binary, Count and Range type suggest that models are usually able to follow instruction, but tend to diverge while generating the counterfactual names, often relying on internal knowledge or producing hallucinated responses, despite explicit instruction to avoid this behavior. In the case of imaginary names, the open source models attain scores close to 0, indicating their inability to generate counterfactual names. Upon reviewing the responses, it was evident that while these models initiate a reasoning, they almost always hallucinate when generating the counterfactual state names. We also see a drop in questions with jumbled values, emphasizing the correlation between values and their corresponding states.

8 Related Work

Visual Question Answering (VQA) has attracted significant attention in computer vision and natural language processing due to its interdisciplinary challenges, as explored by Antol et al. (2015); Goyal et al. (2017); Bazi et al. (2023); Hartsock and Rasool (2024); Zhang et al. (2024). The introduction of Visual Question Rewriting (VQR) by Wei et al. (2021) has further advanced our understanding of how visual information can enhance question-answering systems. Similarly, Wu (2023) introduced visual quizzing, which involves reasoning with both images and their related questions.

Map Question Answering (MQA) and **Chart Question Answering (CQA)** have also emerged as challenging extensions of VQA, requiring the interpretation of visual data representations such as charts and maps. Datasets like ChartQA (Masry et al., 2022) focus on interpreting structured data charts, while Chang et al. (2022) introduced MapQA for choropleth map question answering, highlighting the need for robust VQA systems. MapQA's U.S. focus study and template questions limit its scope. Our dataset on the other hand includes a diverse set of countries, map types and complex questions which were manually curated to create an effective benchmark. Additional details present in the Appendix D.

Enhancing Visual Question Answering. Despite these advances, gaps remain in Chart (CQA) and Map Question Answering (MQA), particularly in handling complex reasoning, numeric answers, and out-of-vocabulary terms. Existing systems often struggle with these challenges, and synthetic datasets may limit their real-world applicability (Bhaisaheb et al., 2023; Chaudhry et al., 2020). Our research addresses these issues by building on Chang et al. (2022) with more diverse maps, challenging questions, and benchmarking state-of-the-art multimodal and visual-language models.

9 Conclusion and Future Work

This paper introduces **MAPWise**, a new large-scale dataset tailored for understanding choropleth maps in three diverse countries: the United States, China, and India. Looking ahead, there are many promising areas for further research based on what we found and from the existing studies. Future studies could broaden the scope of datasets by including different types of maps. Inspired by previous work (Fan et al., 2024), we could complement our dataset by exploring fictional maps or more detailed maps that include features such as rivers and roads. This expansion would help evaluate how well VLMs generalize across diverse geographical contexts. Further research is needed to identify and mitigate biases inherent in map interpretation. Techniques like dataset perturbation, which introduces variations in map features and contexts, could provide deeper insights and help mitigate biases effectively.

To improve how data is extracted, integrating external knowledge sources in future would be a promising strategy. Models that use knowledge graphs, like RAG networks filled with detailed information about state borders and regional relationships, could also improve how well Vision Language Models (VLMs) reason through map-based tasks. Another future direction would be improving how VLMs are trained to recognize colors more accurately and integrating additional datasets, training on auxiliary data such as charts, to improve their ability to interpret and process map-related information effectively.

Future work will also focus on expanding our dataset to overcome its size and diversity limitations. This includes automating question generation using LLMs and data tables, followed by rigorous human verification to maintain quality. We also plan to increase geographic diversity and incorporate a wider variety of map types beyond Choropleth maps.

Limitations

While our study has yielded interesting observations, it's crucial to acknowledge its limitations. We focused exclusively on choropleth maps, which represent data using color gradients. While these maps are effective for visualizing regional data, they lack the detailed features and interactive elements found in more advanced mapping systems like Google Maps.

Moreover, we were limited to maps from only three countries, and the manual question creation process restricted the size of our dataset. Additional work on these aspects to extend the dataset through the addition of more high quality questions, and incorporating more diversity in the types of maps used and the countries or regions they represent would be extremely helpful in expanding the domain further.

Ethics Statement

This research adheres to the ACL code of ethics, acknowledging and addressing potential ethical implications. While LLMs assisted in writing and presentation, all ideas and conclusions are solely attributed to the authors. The research promotes responsible and fair use of methodologies, ensuring transparency and reproducibility. We plan to release all scripts, resources, comprehensive documentation, evaluation metrics, datasets, model specifications, and prompting methods to enable others to build upon our work. We strive to present our findings clearly and accurately, avoiding exaggerated claims or misinterpretations.

Acknowledgment

We are grateful to Arqam Patel, Nirupama Ratna, and Jay Gala for their help with the creation of the MAPWise dataset. Their early efforts were instrumental in guiding the development of this research. We also thank Adnan Qidwai and Jennifer Sheffield for their valuable insights, which helped improve our work. We extend our sincere thanks to the reviewers for their insightful comments and suggestions, which have greatly enhanced the quality of this manuscript.

This research was partially supported by ONR Contract N00014-23-1-2364, and sponsored by the Army Research Office under Grant Number W911NF-20-1-0080. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond.

Yakoub Bazi, Mohamad Mahmoud Al Rahhal, Laila Bashmal, and Mansour Zuair. 2023. Vision–language model for visual question answering in medical imagery. *Bioengineering*, 10(3):380.

S. Bhaisaheb, S. Paliwal, R. Patil, M. Patwardhan, L. Vig, and G. Shroff. 2023. Program synthesis for complex qa on charts via probabilistic grammar based filtered iterative back-translation. *Findings of the Association for Computational Linguistics: EACL 2023*.

Fengxiang Bie, Yibo Yang, Zhongzhu Zhou, Adam Ghanem, Minjia Zhang, Zhewei Yao, Xiaoxia Wu, Connor Holmes, Pareesa Ameneh Golnari, David A. Clifton, Yuxiong He, Dacheng Tao, and Shuaiwen Leon Song. 2023. Renaissance: A survey into ai text-to-image generation in the era of large model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47:2212–2231.

Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. 2022. Mapqa: A dataset for question answering on choropleth maps. In *Table Representation Learning Workshop, NeurIPS Workshops*.

Ritwick Chaudhry, Sumit Shekhar, Utkarsh Gupta, Pranav Maneriker, Prann Bansal, and Ajay Joshi. 2020. Leaf-qa: Locate, encode & attend for figure question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3512–3521.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *Preprint*, arXiv:2401.16420.

Arlen Fan, Fan Lei, Michelle Mancenido, Alan M Maceachren, and Ross Maciejewski. 2024. Understanding reader takeaways in thematic maps under varying text, detail, and spatial autocorrelation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Team Gemini. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Iryna Hartsock and Ghulam Rasool. 2024. Vision-language models for medical report generation and visual question answering: A review. *Frontiers in Artificial Intelligence*, 7:1430984.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. 2024. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14281–14290.

Team InternLM2. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Raihan Kabir, Naznin Haque, Md Saiful Islam, et al. 2024. A comprehensive survey on visual question answering datasets and algorithms. *arXiv preprint arXiv:2411.11150*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Songning Lai, Xifeng Hu, Haoxuan Xu, Zhaoxia Ren, and Zhi Liu. 2023. Multimodal sentiment analysis: A survey. *Displays*, page 102563.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Team Qwen. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

J. Wei, X. Li, Y. Zhang, and X. Wang. 2021. Visual question rewriting for increasing response rate. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

K. Wu. 2023. Research and implementation of visual question and answer system based on deep learning. *Applied Mathematics and Nonlinear Sciences*, 9.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Appendix

A Remaining Counter Factual Results

In this section we display the result for the remaining open-source models for zero-shot COT prompt (Table 12) and results for all models for the EER prompt (Table 13) from the study.

CF Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall
CogAgent-VQA						
Original	36.73	19.79	4.44	10.42	20.39	40.14
Imaginary	55.10	0.00	17.78	10.42	3.06	3.06
Shuffled	57.14	17.71	8.89	5.21	7.73	9.52
Jumbled	34.69	15.63	4.44	6.25	30.35	43.54
QwenVL						
Original	48.98	8.33	15.56	5.21	15.84	30.78
Imaginary	47.92	2.08	13.33	9.57	3.01	10.88
Shuffled	51.02	6.25	11.11	13.54	6.95	12.24
Jumbled	38.78	10.42	4.44	10.42	25.24	38.78

Table 12: Counter Factual Results (in %) for zero-shot COT prompt for CogAgent and QwenVL. CF represents Counter Factual

B Dataset Creation and Validation

In our study, we engaged a total of 6 annotators from our research group. Given their expertise and familiarity with our project goals, these individuals voluntarily contributed their time and knowledge without financial compensation. We believed that their intrinsic motivation to improve NLP research and their commitment to the project's objectives outweighed the need for monetary incentives.

Following the initial annotation process, the ground truth answers were established through a rigorous verification process with the help of two additional annotators to ensure accuracy and minimize subjectivity. For region-based question, we adhered to widely accepted geographical definitions and cross referenced them with readily available online resources.

CF Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall
Gemini 1.5 Flash						
Original	68.75	39.58	26.67	50.00	43.25	63.27
Imaginary	61.22	43.75	20.00	41.67	24.91	36.73
Shuffled	71.43	25.00	13.33	37.76	6.80	9.52
Jumbled	66.67	28.13	13.33	48.96	43.58	41.50
GPT-4o						
Original	70.00	42.86	28.89	53.33	45.68	65.21
Imaginary	63.21	45.83	22.22	43.75	26.54	38.29
Shuffled	70.00	24.17	12.22	36.46	6.12	8.75
Jumbled	68.75	30.56	15.56	51.04	46.72	43.87
Idefics2						
Original	53.06	23.96	4.44	25.00	27.66	61.39
Imaginary	53.06	0.00	22.22	16.67	0.19	1.70
Shuffled	51.02	13.54	26.67	21.88	11.12	21.09
Jumbled	51.02	13.54	8.89	32.29	24.07	47.62
InternLM-XComposer2						
Original	53.06	11.46	13.33	17.71	35.19	48.64
Imaginary	51.02	0.00	17.78	7.29	3.32	7.14
Shuffled	53.06	12.50	15.56	14.58	8.10	15.99
Jumbled	34.69	8.33	8.89	8.33	23.38	37.59
CogAgent-VQA						
Original	44.90	25.00	15.56	15.63	20.48	38.44
Imaginary	42.86	0.00	13.33	12.50	2.70	3.74
Shuffled	51.02	20.83	22.22	23.96	10.44	12.59
Jumbled	40.82	21.88	17.78	21.88	15.27	29.93
QwenVL						
Original	46.94	14.58	13.33	4.17	13.33	26.87
Imaginary	51.02	0.00	8.89	6.25	1.08	5.61
Shuffled	53.06	9.38	17.78	6.25	5.58	13.61
Jumbled	46.94	18.75	17.78	15.63	13.39	23.81

Table 13: Counter Factual Results (in %) for EER prompt for all models in the study. CF represents Counter Factual and Accuracy stands for Accuracy

Initial annotation took approximately one minute on average, with more time required for questions involving spatial reasoning or external knowledge about the geographic regions of a country. The verification process was less time consuming, with each question taking around 20 to 30 seconds on average. The entire annotation and verification process took approximately four weeks.

C Rank Wise Precision (RWP) Vs MAP and MRR

The main purpose of introducing the Rank Wise Precision (RWP) score (Algorithm 1 for computing RWP score) was to avoid giving different scores based on the order of the states within the same rank. Traditional metrics such as Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP) assign higher scores to states that appear first in the order. However, for our evaluation, we are concerned with the states irrespective of their order within the same rank. For example, consider the ground truth ranks as follows

- Rank 1: [California]
- Rank 2: [Washington]
- Rank 3: [Oregon]

When the model is asked to rank the states based on a range value according to the color or shape on a map, it first identifies the color or shape. If more than one state has the same color, they are give the same rank. Consider the two cases:

- Case 1: The model's output is Rank 1: [California], Rank 2 : [Washington, Oregon]
- Case 2: The model's output is Rank 1: [California], Rank 2: [Oregon, Washington]

Algorithm 1 Calculate Rank Wise Precision (RWP) Score

```
Initialize an empty list RWP
for each rank in ground_truth_ranks do
    g_items ← items in the ground_truth for the current rank
    p_items ← items in predicted order for the current rank
    Append precision(g_items, p_items) to RWP
end for
return mean(RWP)
```

In both cases, all three metrics will give a score of 1 for Rank 1 and a score of 0 for Rank 3. However, for Rank 2, MRR will give a score of 1 for Case 1 and 0.5 for Case 2. MAP will give a score of 0.75 for Case 1 and 0.25 for Case 2. In contrast, RWP will give a score of 0.5 for both cases. Therefore, RWP scores are agnostic to the order of states within the same rank, for final score we take the mean of the scores of all 3 ranks. (Table 14 and 15 shows the RWP, MAP and MRR scores for India, China and USA).

D Comparison with MapQA dataset

While MapQA is a valuable resource with its large dataset of 800,000 question-answer pairs, our work distinguishes itself by addressing crucial limitations in MapQA's scope and analytical depth.

Targeted Dataset Design and Complexity:

- Our dataset, while smaller in scale than MapQA (3,000 question-answer pairs), is meticulously curated to specifically test complex reasoning skills related to choropleth maps.
- We focus on challenging aspects of choropleth map interpretation, ensuring high-quality data for precise model evaluation.

Map Type	India			China			USA		
	MRR	MAP	RWP	MRR	MAP	RWP	MRR	MAP	RWP
GPT-4o									
With Annotations	57.41%	54.94%	53.70%	48.40%	45.66%	46.58%	53.33%	48.19%	49.17%
Without Annotations	57.41%	55.25%	55.56%	51.21%	50.52%	50.48%	54.72%	51.51%	52.04%
Hactched	53.57%	52.98%	53.57%	39.58%	39.58%	39.58%	62.96%	59.26%	60.19%
Gemini 1.5 Flash									
With Annotations	40.48%	38.99%	38.69%	57.80%	54.31%	53.63%	57.13%	52.61%	53.80%
Without Annotations	49.40%	46.73%	45.83%	42.95%	41.35%	41.03%	49.72%	44.03%	43.61%
Hactched	38.89%	38.33%	38.89%	61.46%	60.94%	60.42%	47.22%	43.52%	43.52%
Idefics2									
With Annotations	45.24%	45.24%	45.24%	46.15%	46.15%	46.15%	42.22%	42.22%	42.22%
Without Annotations	48.81%	48.81%	48.81%	49.36%	49.36%	49.36%	40.37%	40.12%	40.37%
Hactched	31.11%	31.11%	31.11%	34.38%	34.38%	34.38%	45.37%	45.37%	45.37%
InternLM-XComposer2									
With Annotations	35.71%	35.71%	35.71%	42.31%	41.99%	41.67%	42.22%	41.94%	42.22%
Without Annotations	39.29%	39.29%	39.29%	39.74%	39.74%	39.74%	38.33%	38.33%	38.33%
Hactched	44.44%	44.44%	44.44%	47.92%	47.92%	47.92%	36.11%	36.11%	36.11%

Table 14: Comparing our RWP score with other popular MRR and MAP rank scores for zero-shot COT prompt

Map Type	India			China			USA		
	MRR	MAP	RWP	MRR	MAP	RWP	MRR	MAP	RWP
GPT-4o									
With Annotations	64.20%	62.07%	61.52%	59.08%	56.59%	56.62%	52.78%	46.94%	48.33%
Without Annotations	56.17%	53.40%	53.09%	62.50%	60.90%	60.26%	46.67%	41.14%	40.93%
Hactched	64.29%	63.10%	61.90%	39.44%	38.01%	38.15%	60.19%	54.63%	55.56%
Gemini 1.5 Flash									
With Annotations	38.27%	36.15%	35.60%	53.82%	50.54%	49.54%	57.41%	50.63%	50.65%
Without Annotations	61.86%	60.26%	60.26%	41.99%	37.18%	35.90%	51.30%	44.41%	44.63%
Hactched	45.24%	44.05%	42.86%	38.54%	35.94%	35.42%	41.98%	32.79%	32.87%
Idefics2									
With Annotations	48.81%	48.81%	48.81%	28.85%	28.53%	28.21%	40.56%	40.28%	40.00%
Without Annotations	39.29%	39.29%	39.29%	37.18%	37.18%	37.18%	36.67%	36.67%	36.67%
Hactched	26.67%	26.67%	26.67%	39.58%	39.58%	39.58%	42.59%	42.59%	42.59%
InternLM-XComposer2									
With Annotations	53.57%	53.57%	53.57%	41.67%	41.67%	41.67%	41.67%	41.39%	41.11%
Without Annotations	47.62%	47.62%	47.62%	46.15%	46.15%	46.15%	36.11%	36.11%	36.11%
Hactched	44.44%	44.44%	44.44%	35.42%	35.42%	35.42%	45.37%	45.37%	45.37%

Table 15: Comparing our RWP score with other popular MRR and MAP rank scores for EER prompt

- We incorporate a variety of map types, including continuous and discrete maps with diverse visual representations, such as variations in legend placement, background presence, and colormaps. Additionally, we include real-world map types like hatched maps, increasing the task's complexity.
- We analyze both annotated and unannotated maps to further understand how different map types influence question answering performance.
- Unlike MapQA's automatically generated questions, our human-annotated questions require nuanced understanding of relative spatial relationships, intricate map features, and complex reasoning, moving beyond simple information retrieval.

For instance, our dataset includes questions such as: "Which two regions that are closest to each other belong to the largest range?" Answering this question necessitates not only identifying the largest range but also using data extraction techniques to find regions within that range. Moreover, models need to rely on visual cues from the map and their internal knowledge base to correctly identify regions that satisfy both the range criteria and proximity requirements.

Another complex example from our dataset is: "Name the southernmost state that belongs to a higher value range compared to all its neighbors." To answer this, models must extract value data for each state, compare those values with their neighbors, and then utilize visual data or internal knowledge to identify the southernmost state among those meeting the criteria.

Additional Diverse Domains:

- MapQA is limited to maps of the USA, whereas our dataset includes maps from three countries (USA, India, and China), helping to highlight potential biases in model understanding of diverse regions.

Advanced Analysis and Novel Contributions:

- Our analysis surpasses MapQA's scope by encompassing a broader range of models, including open and closed-source Vision-Language Models (VLMs) and Multimodal Language Models (MLLMs). This comprehensive evaluation provides a more accurate picture of the current state-of-the-art in choropleth map understanding and identifies promising avenues for future research.

- We go beyond overall accuracy metrics by providing a detailed breakdown of model performance across different answer types. This granular analysis, missing in MapQA, pinpoints areas where models struggle, guiding future research towards targeted improvements in choropleth map understanding.

- By evaluating model performance on data with imaginary state names, jumbled state names, and synthetic information, we offer critical insights into model robustness and generalization, pushing the boundaries of current evaluation methods.

In conclusion, while MapQA establishes a strong foundation for map-based question answering, our work delves deeper into the complexities of choropleth maps. Our meticulously designed dataset, novel counterfactual analysis, and comprehensive model evaluation provide a more challenging benchmark and a nuanced understanding of model capabilities, paving the way for further advancements in this crucial field.

E Zero Shot - CoT Prompt

The prompt we used for analysis using zero shot COT has been presented in Figure 4.

F Few Shot - CoT

In addition to Zero Shot COT, we also tried Few shot COT. In this approach, we included several examples within the prompt, anticipating that the model would adopt the demonstrated reasoning style before providing its final answer. Given that the task involves both textual and visual modalities, it is crucial to provide different visual cues for the examples to prevent hallucinations caused by manual intervention. We addressed this issue using two sub-approaches:

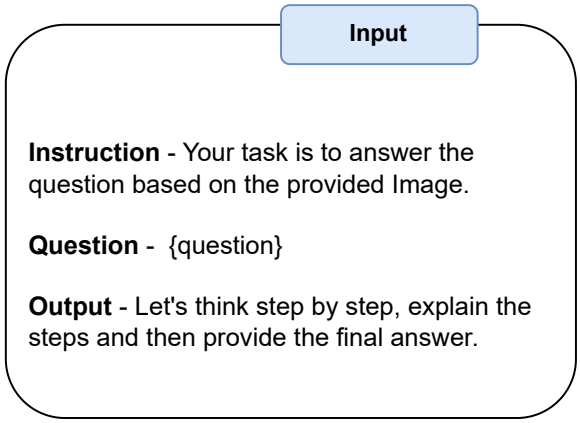


Figure 4: Zero shot COT prompt representation

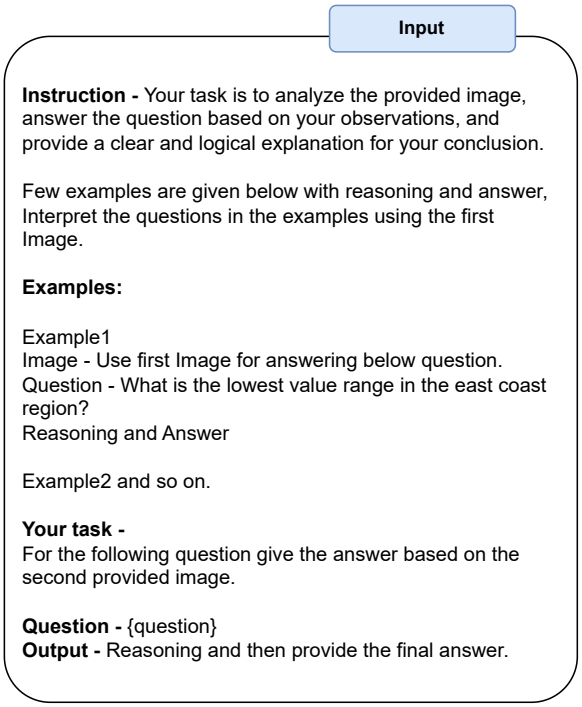


Figure 5: Example of a Few shot COT with second image for example

- **Textual Conversion of Visual Representation**: The visual map corresponding to example was converted into textual description. (see Figure 6 for the prompt style)

- **Inclusion of a Second Image in the Prompt**: In this, we provided a separate image for the examples. (see Figure 5 for the prompt style)

Input

Instruction - Your task is to analyze the provided image, answer the question based on your observations, and provide a clear and logical explanation for your conclusion.

Few examples are given below with reasoning and answer, Interpret the questions in the examples using the Image explanation below Examples.

Examples:

Image Explanation only for examples -:
The image is a Choropleth map of Australia that covers state and territories.
The map uses different shades of blue to represent different ranges, and the colors are as follows:

Very Light Blue: 27 - 136
Light Blue: 136 - 482.5
Medium Blue: 482.5 - 1,149
Dark Blue: 1,149 - 3,075

States with Dark Blue color -
New South Wales, Victoria

so on

end of image explanation
###

Example1
Image - Use above Image explanation for answering the below question.
Question - What is the lowest value range in the east coast region?
Reasoning and Answer

Example2 and so on.

Your task -
For the following question give the answer based on the provided image.

Question - {question}

Output - Reasoning and then provide the final answer.

Figure 6: Example of a Few shot COT with visual to textual representation.

To avoid introducing any unintended bias through the examples, we prepared examples involving a country not represented in the MAPWise Dataset (Table 16 represents the Few shot results). Largely, Few shots with textual conversion of visual representation (VTM) works better for all map types and country.

G EER Prompt

We implemented an explicit reasoning prompt, EER which focused on extracting the vital information from models step by step before arriving at the final answer. The prompt used for the same is presented in Figure 7

Input

Instruction - You are an expert at answering questions based on maps. You will be given a map and a question and you will have to answer the question through the following four steps.

Make sure to answer the question step by step strictly following the 4 steps mentioned and answering only on the basis of the map provided.

Step 1: Extract the names of the regions which are relevant for answering the given question.

For the given question, you will have to extract the names of all the relevant regions which are required for answering the question. If the question mentions a state directly, also consider that in the list of relevant regions.

Here are a few examples:
...

Your answer should be in the format of a list with the names of all the relevant regions.

Step 2: Extract the names of the states in the relevant regions

For each relevant region, extract the names of the states which fall in the region and would be required for answering the questions.

Here are a few examples:
...

Step 3: Extract the values corresponding to the states from the given map.

For all the relevant states extracted in step 2, you will have to refer to the map and extract the corresponding values for those states from the map by using the legend given in the map.
Do this for each state individually. You can treat this as a task for simply extracting values from a map according to the legend.

For each relevant state "X" in the list, answer the question "What is the value for state X in the given map?"

Report all answers at the end in the form of a table.
Step 4: Answering the question based on the extracted data.

To answer the question you have to use the table of states and their corresponding values as extracted in step 3. For answering the question, you should use reasoning and think step by step to arrive at the final answer. State all your reasoning steps.

Your Taks-
Follow the above steps to answer the following question based on the given map.

Question - {question}

Steps:

Figure 7: Example of EER Prompt

Prompt	Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall	Rank RWP
				USA				
VTM	With Annotations	68.30%	61.41%	55.10%	66.21%	38.23%	46.94%	42.04%
	Without Annotations	67.26%	63.91%	55.10%	68.28%	33.96%	38.87%	49.54%
	Hatched	54.08%	57.67%	51.22%	48.68%	27.11%	40.65%	43.06%
SIE	With Annotations	64.29%	50.61%	46.94%	59.66%	31.99%	37.61%	37.87%
	Without Annotations	60.13%	55.00%	53.06%	62.76%	28.66%	34.86%	44.26%
	Hatched	50.21%	56.98%	40.00%	47.37%	30.19%	37.54%	37.81%
				India				
VTM	With Annotations	65.35%	45.81%	23.40%	45.58%	36.72%	42.91%	38.70%
	Without Annotations	61.62%	43.80%	29.79%	47.35%	37.28%	43.75%	35.63%
	Hatched	57.14%	48.65%	29.41%	40.63%	30.67%	33.58%	40.00%
SIE	With Annotations	58.33%	42.31%	34.04%	41.59%	38.03%	43.06%	46.55%
	Without Annotations	56.14%	46.79%	27.66%	42.92%	40.19%	46.83%	33.14%
	Hatched	58.79%	45.27%	27.45%	25.00%	33.63%	43.31%	46.67%
				China				
VTM	With Annotations	60.00%	56.02%	14.43%	52.97%	29.37%	36.15%	41.77%
	Without Annotations	64.55%	58.60%	17.53%	56.78%	34.56%	44.79%	49.47%
	Hatched	52.72%	47.19%	29.63%	38.71%	30.52%	44.44%	29.17%
SIE	With Annotations	64.77%	52.26%	14.43%	47.03%	33.10%	41.92%	31.84%
	Without Annotations	63.18%	54.95%	19.59%	49.15%	29.37%	39.26%	40.71%
	Hatched	51.63%	49.57%	29.63%	30.65%	28.43%	39.32%	32.81%

Table 16: Chain of Thought with Few shot results (in %) for Gemini model. VTM stands for (visual to textual modality) and SIE stand for (separate image for examples)

H Comprehensive Results

In this section, we present the complete results for two prompts - Zero shot COT and Explicit Extraction and Reasoning (EER) - across all countries, map types and models. This comprehensive coverage provides a detailed comparison of the performance variations under different conditions.

Answer Type	USA	India	China
Count	11	17	30
Range	1	8	2
Binary	84	121	120
Single	32	30	33
List	17	39	58
Rank	0	0	0

Table 17: Counts for country-wise relative region based questions from each category of expected answer type

For questions explicitly requiring the use of relative regions, the distribution of answer types for each question is provided in Table 17. This offers better insight into the results presented in that section.

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall	Rank RWP
GPT-4o							
With Annotations	70.26	44.68	51.02	71.28	38.64	45.61	49.17%
Without Annotations	68.85	39.38	53.06	65.71	41.58	47.02	52.04%
Hactched	49.78	16.14	26.83	26.67	26.96	32.60	60.19%
Gemini 1.5 Flash							
With Annotations	65.03	54.65	51.02	63.10	38.73	51.43	53.80%
Without Annotations	65.70	59.55	42.86	64.48	33.26	44.04	43.61%
Hactched	49.36	56.20	51.22	53.95	20.51	35.35	43.52%
Idefics2							
With Annotations	54.79	61.11	18.37	29.66	26.64	42.99	42.22%
Without Annotations	55.23	59.09	24.49	30.69	29.61	43.72	40.37%
Hactched	52.79	56.98	12.20	25.00	22.99	42.09	45.37%
InternLM-XComposer2							
With Annotations	52.78	35.86	32.65	20.00	22.63	33.07	42.22%
Without Annotations	52.78	42.93	44.90	22.07	20.81	30.12	38.33%
Hactched	45.49	31.40	53.66	17.11	25.84	31.31	36.11%
CogAgent-VQA							
With Annotations	44.03	42.23	24.49	23.40	19.79	33.00	41.67%
Without Annotations	39.34	42.23	22.45	25.18	20.16	27.31	43.89%
Hactched	34.67	41.34	24.39	20.00	21.66	25.09	41.67%
QwenVL							
With Annotations	37.72	20.19	3.19	6.82	19.24	28.59	33.33%
Without Annotations	35.09	16.67	5.32	8.18	18.35	28.48	35.56%
Hactched	32.42	9.46	3.92	2.46	18.42	24.32	33.33%

Table 18: USA results for all models in the study with zero-shot COT prompt

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall	Rank RWP
GPT-4o							
With Annotations	65.57	62.02	51.02	60.64	37.78	45.21	48.33%
Without Annotations	62.30	59.60	59.18	60.99	39.00	44.82	40.93%
Hactched	36.89	30.71	21.95	28.00	17.48	24.36	55.56%
Gemini 1.5 Flash							
With Annotations	66.74	61.19	51.02	62.41	42.71	50.28	50.65%
Without Annotations	64.17	62.28	40.82	61.70	37.32	45.05	44.63%
Hactched	48.44	40.55	39.02	40.00	18.80	34.62	32.87%
Idefics2							
With Annotations	56.57	27.02	6.12	28.97	23.80	46.41	40.00%
Without Annotations	53.45	26.26	12.24	26.21	21.58	45.83	36.67%
Hactched	56.65	25.19	14.63	19.74	21.24	44.87	42.59%
InternLM-XComposer2							
With Annotations	49.67	22.22	16.33	20.69	24.41	35.13	41.11%
Without Annotations	47.66	33.08	26.53	19.66	25.74	36.71	36.11%
Hactched	47.21	32.17	31.71	11.84	17.72	26.60	45.37%
CogAgent-VQA							
With Annotations	28.10	20.21	0.00	17.38	19.75	31.08	43.33%
Without Annotations	26.70	30.05	14.29	16.67	20.99	27.36	42.04%
Hactched	28.89	26.77	9.76	10.67	11.15	15.93	49.07%
QwenVL							
With Annotations	23.46	16.67	6.38	5.45	19.88	32.38	27.78%
Without Annotations	29.82	9.62	3.19	9.09	21.28	30.22	23.89%
Hactched	29.67	14.19	5.88	3.28	19.94	26.18	20.37%

Table 19: USA results for all models in the study with EER prompt

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall	Rank RWP
GPT-4o							
With Annotations	71.52	40.06	35.48	55.75	49.94	49.17	54.94
Without Annotations	66.45	40.92	30.85	53.54	46.23	47.09	55.25
Hactched	49.72	28.38	31.37	30.77	34.19	36.27	52.98
Gemini 1.5 Flash							
With Annotations	56.36	38.49	24.47	40.27	34.55	45.11	38.99
Without Annotations	58.99	37.39	23.40	35.84	36.25	46.21	46.73
Hactched	52.75	48.65	23.53	34.38	38.95	47.33	38.33
Idefics2							
With Annotations	54.17	43.59	13.83	28.76	32.19	38.29	45.24
Without Annotations	56.58	43.48	15.96	23.45	28.04	36.26	48.81
Hactched	47.80	45.95	17.65	20.31	23.31	30.64	31.11
InternLM-XComposer2							
With Annotations	56.80	32.37	17.02	13.27	20.14	24.02	35.71
Without Annotations	53.51	34.08	14.89	13.27	27.84	35.84	39.29
Hactched	51.65	24.32	21.57	24.74	29.83	44.44	44.44
CogAgent-VQA							
With Annotations	43.27	25.32	9.57	16.81	19.62	26.32	41.36
Without Annotations	44.37	29.70	10.64	15.49	22.64	28.86	38.89
Hactched	43.09	29.50	7.84	4.69	17.30	20.32	46.69
QwenVL							
With Annotations	37.75	22.33	4.26	6.64	17.00	23.60	17.31
Without Annotations	35.10	16.67	5.32	8.85	15.50	19.55	35.19
Hactched	32.04	9.46	5.88	2.34	19.09	22.55	21.43

Table 20: India results for all models in the study with zero-shot COT prompt

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall	Rank RWP
GPT-4o							
With Annotations	65.12	52.46	40.00	52.23	49.88	51.51	62.07
Without Annotations	65.34	51.82	40.43	45.13	46.98	47.30	53.40
Hactched	43.65	39.86	33.33	28.13	29.46	30.76	63.10
Gemini 1.5 Flash							
With Annotations	61.37	37.82	23.40	39.38	29.01	37.32	36.15
Without Annotations	62.69	42.52	24.47	43.81	39.13	51.44	60.26
Hactched	58.01	40.32	33.33	31.25	38.01	53.24	44.05
Idefics2							
With Annotations	55.26	36.54	8.51	21.24	32.65	48.75	48.81
Without Annotations	52.41	38.25	11.70	21.68	28.17	46.45	39.29
Hactched	54.40	29.73	5.88	12.50	26.31	37.28	26.67
InternLM-XComposer2							
With Annotations	51.32	25.96	15.96	13.72	24.20	27.80	53.57
Without Annotations	52.41	28.85	8.51	13.72	25.83	26.68	47.62
Hactched	48.35	13.51	17.65	1.56	17.70	17.28	44.44
CogAgent-VQA							
With Annotations	30.46	10.90	4.26	8.41	16.55	19.88	40.12
Without Annotations	33.33	13.46	4.26	9.29	17.52	19.84	42.42
Hactched	31.49	9.46	0.00	6.25	15.51	17.84	29.23
QwenVL							
With Annotations	23.18	17.95	5.32	5.31	11.64	20.43	29.01
Without Annotations	30.02	9.62	1.06	8.85	16.61	23.10	32.10
Hactched	29.83	14.19	3.92	3.13	17.13	21.81	16.67

Table 21: India results for all models in the study with EER prompt

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall	Rank RWP
GPT-4o							
With Annotations	66.97	47.53	50.52	59.40	53.93	57.56	45.66
Without Annotations	68.10	41.61	43.18	61.27	44.17	47.32	50.52
Hactched	45.11	20.56	27.78	18.03	33.33	34.40	39.58
Gemini 1.5 Flash							
With Annotations	62.27	51.83	13.40	52.97	22.76	38.96	54.31
Without Annotations	60.23	51.83	13.40	53.81	31.56	44.84	41.35
Hactched	53.80	55.41	22.22	40.32	29.61	45.41	60.94
Idefics2							
With Annotations	54.09	50.54	21.65	34.32	21.67	29.91	46.15
Without Annotations	56.82	48.17	19.59	30.93	21.86	28.73	49.36
Hactched	56.52	47.62	16.67	14.52	23.01	33.33	34.38
InternLM-XComposer2							
With Annotations	54.09	38.39	19.59	28.81	22.98	28.02	41.99
Without Annotations	53.86	38.49	22.68	26.27	27.28	33.84	39.74
Hactched	50.54	32.90	22.22	4.84	27.37	31.62	47.92
CogAgent-VQA							
With Annotations	48.64	28.39	20.62	22.03	21.32	27.91	48.08
Without Annotations	46.47	28.60	11.34	30.34	29.08	31.83	26.92
Hactched	47.28	42.64	11.11	9.84	26.39	32.05	31.48
QwenVL							
With Annotations	37.81	21.18	3.09	9.40	23.92	27.90	21.79
Without Annotations	36.90	23.66	2.06	18.38	22.56	27.69	19.23
Hactched	40.76	17.32	11.11	3.28	23.34	27.35	12.50

Table 22: China results for all models in the study with zero-shot COT prompt

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall	Rank RWP
GPT-4o							
With Annotations	63.33	60.65	45.36	59.83	43.47	46.48	56.59
Without Annotations	64.01	58.39	49.48	54.27	47.24	50.93	60.90
Hactched	44.02	31.39	24.07	21.31	33.01	37.82	38.01
Gemini 1.5 Flash							
With Annotations	61.50	54.09	24.74	52.14	23.01	39.54	50.54
Without Annotations	59.23	53.23	28.87	53.42	29.06	41.46	37.18
Hactched	47.83	48.92	25.93	36.07	30.84	50.43	35.94
Idefics2							
With Annotations	53.86	29.25	18.56	28.39	26.63	39.78	28.53
Without Annotations	55.00	29.78	19.59	26.69	25.70	37.12	37.18
Hactched	52.17	18.18	12.96	9.68	34.03	47.44	39.58
InternLM-XComposer2							
With Annotations	42.50	23.66	21.65	24.15	20.97	24.74	41.67
Without Annotations	45.91	32.04	21.65	19.92	24.83	27.86	46.15
Hactched	48.37	20.78	18.52	2.42	30.45	33.65	35.42
CogAgent-VQA							
With Annotations	27.56	19.35	5.15	18.38	22.25	25.62	25.46
Without Annotations	24.15	16.77	9.28	18.38	27.61	30.12	32.05
Hactched	33.70	22.51	0.00	8.20	31.40	35.58	14.58
QwenVL							
With Annotations	21.18	8.92	2.06	8.12	24.01	29.35	24.36
Without Annotations	24.37	13.44	2.06	8.55	22.81	25.98	16.67
Hactched	26.09	7.58	1.85	6.56	20.29	26.71	14.58

Table 23: China results for all models in the study with EER prompt

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall	Rank RWP
GPT-4o							
With Annotations	64.05%	62.12%	25.00%	61.84%	36.70%	42.11%	56.94%
Without Annotations	67.97%	71.21%	25.00%	55.26%	43.96%	46.78%	62.50%
Gemini 1.5 Flash							
With Annotations	63.03%	56.52%	25.00%	43.75%	38.84%	53.39%	43.98%
Without Annotations	60.61%	62.32%	25.00%	46.25%	31.37%	43.22%	44.44%
Idefics2							
With Annotations	61.21%	63.77%	12.50%	27.50%	23.55%	41.24%	30.56%
Without Annotations	61.82%	60.87%	50.00%	36.25%	29.84%	41.81%	27.78%
InternLM-XComposer2							
With Annotations	54.55%	50.72%	12.50%	22.50%	23.56%	34.18%	33.33%
Without Annotations	53.94%	59.42%	37.50%	25.00%	22.40%	33.90%	36.11%
CogAgent-VQA							
With Annotations	44.44%	16.67%	12.50%	31.58%	20.69%	30.41%	33.33%
Without Annotations	39.22%	13.64%	12.50%	27.63%	25.00%	32.46%	33.33%
QwenVL							
With Annotations	35.71%	24.39%	2.33%	8.33%	19.56%	29.94%	19.44%
Without Annotations	33.33%	19.51%	4.65%	15.28%	20.59%	31.36%	27.78%

Table 24: USA results for all models in the study with zero-shot COT prompt for continuous maps only

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall	Rank RWP
GPT-4o							
With Annotations	64.71%	69.70%	12.50%	50.00%	42.12%	44.74%	52.78%
Without Annotations	64.05%	66.67%	50.00%	47.37%	47.25%	48.83%	52.78%
Gemini 1.5 Flash							
With Annotations	62.09%	72.73%	12.50%	44.74%	38.96%	46.78%	37.50%
Without Annotations	64.71%	69.70%	25.00%	44.74%	36.55%	44.44%	42.59%
Idefics2							
With Annotations	60.61%	20.29%	12.50%	35.00%	27.78%	48.02%	29.17%
Without Annotations	59.39%	20.29%	25.00%	32.50%	24.56%	46.61%	22.22%
InternLM-XComposer2							
With Annotations	53.94%	24.64%	12.50%	22.50%	28.63%	39.55%	41.67%
Without Annotations	51.52%	33.33%	12.50%	26.25%	32.94%	44.07%	38.89%
CogAgent-VQA							
With Annotations	28.10%	27.27%	0.00%	22.37%	24.93%	40.06%	41.67%
Without Annotations	25.49%	16.67%	0.00%	19.74%	26.80%	33.04%	34.26%
QwenVL							
With Annotations	22.22%	14.63%	6.98%	8.33%	21.51%	35.59%	27.78%
Without Annotations	29.76%	10.98%	2.33%	16.67%	20.54%	33.05%	25.00%

Table 25: USA results for all models in the study with EER prompt for continuous maps only

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall	Rank RWP
GPT-4o							
With Annotations	72.00%	40.24%	26.19%	58.33%	54.07%	53.29%	65.38%
Without Annotations	70.00%	41.46%	25.58%	51.39%	52.42%	54.47%	71.79%
Gemini 1.5 Flash							
With Annotations	53.57%	29.63%	18.60%	43.06%	37.12%	51.16%	58.97%
Without Annotations	59.13%	28.05%	23.26%	43.06%	37.86%	49.60%	74.36%
Idefics2							
With Annotations	53.97%	51.22%	9.30%	43.06%	31.53%	39.38%	52.56%
Without Annotations	56.35%	47.56%	13.95%	29.17%	31.70%	42.39%	56.41%
InternLM-XComposer2							
With Annotations	60.32%	43.90%	18.60%	22.22%	17.36%	22.42%	38.46%
Without Annotations	54.37%	43.90%	13.95%	22.22%	22.83%	34.42%	43.59%
CogAgent-VQA							
With Annotations	44.40%	14.63%	6.98%	27.78%	18.06%	24.29%	39.74%
Without Annotations	44.00%	24.39%	9.30%	23.61%	17.98%	25.71%	34.62%
QwenVL							
With Annotations	35.60%	25.61%	4.65%	8.33%	17.41%	24.49%	32.05%
Without Annotations	33.60%	19.51%	4.65%	15.28%	15.85%	20.97%	32.05%

Table 26: India results for all models in the study with zero-shot COT prompt for continuous maps only

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall	Rank RWP
GPT-4o							
With Annotations	65.60%	56.10%	30.23%	54.17%	56.10%	56.40%	83.33%
Without Annotations	70.40%	56.10%	34.88%	50.00%	48.63%	49.90%	53.85%
Gemini 1.5 Flash							
With Annotations	58.00%	36.59%	13.95%	51.39%	33.29%	41.46%	43.16%
Without Annotations	63.20%	40.24%	16.28%	48.61%	43.51%	56.81%	83.33%
Idefics2							
With Annotations	61.90%	42.68%	11.63%	27.78%	33.14%	51.85%	56.41%
Without Annotations	58.33%	45.12%	11.63%	37.50%	28.99%	50.07%	35.90%
InternLM-XComposer2							
With Annotations	52.38%	32.93%	18.60%	26.39%	27.30%	32.54%	50.00%
Without Annotations	51.98%	35.37%	4.65%	23.61%	25.04%	28.47%	44.87%
CogAgent-VQA							
With Annotations	31.60%	9.76%	4.65%	15.28%	18.73%	21.24%	44.87%
Without Annotations	34.40%	9.76%	4.65%	15.28%	17.78%	20.26%	37.18%
QwenVL							
With Annotations	21.60%	15.85%	6.98%	8.33%	10.62%	18.90%	30.77%
Without Annotations	30.00%	10.98%	2.33%	16.67%	13.04%	20.46%	28.21%

Table 27: India results for all models in the study with EER prompt for continuous maps only

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall	Rank RWP
GPT-4o							
With Annotations	70.93%	54.49%	51.16%	75.00%	56.93%	60.84%	52.22%
Without Annotations	66.96%	41.03%	39.53%	79.35%	46.39%	49.50%	64.44%
Gemini 1.5 Flash							
With Annotations	62.28%	53.85%	16.28%	66.30%	23.13%	39.31%	38.33%
Without Annotations	62.72%	55.77%	11.63%	70.65%	31.70%	43.53%	45.00%
Idefics2							
With Annotations	50.44%	50.00%	25.58%	57.61%	19.80%	26.08%	43.33%
Without Annotations	51.75%	46.15%	25.58%	53.26%	19.82%	25.69%	56.67%
InternLM-XComposer2							
With Annotations	53.95%	46.15%	18.60%	46.74%	29.17%	32.94%	20.00%
Without Annotations	49.12%	46.15%	30.23%	41.30%	29.79%	36.67%	36.67%
CogAgent-VQA							
With Annotations	45.61%	16.67%	13.95%	41.30%	23.79%	29.41%	50.00%
Without Annotations	42.73%	15.38%	6.98%	46.74%	32.56%	33.13%	33.33%
QwenVL							
With Annotations	34.36%	21.15%	2.33%	15.22%	24.51%	29.02%	23.33%
Without Annotations	33.92%	29.49%	4.65%	27.17%	23.93%	28.51%	16.67%

Table 28: China results for all models in the study with zero-shot COT prompt for continuous maps only

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall	Rank RWP
GPT-4o							
With Annotations	69.16%	68.59%	41.86%	75.00%	45.03%	47.99%	52.78%
Without Annotations	70.04%	67.95%	51.16%	77.17%	50.32%	54.52%	61.67%
Gemini 1.5 Flash							
With Annotations	60.35%	62.82%	18.60%	71.74%	24.51%	39.56%	41.98%
Without Annotations	57.71%	64.10%	20.93%	70.65%	27.40%	40.16%	35.00%
Idefics2							
With Annotations	56.14%	33.33%	18.60%	46.74%	23.81%	32.16%	30.00%
Without Annotations	55.70%	33.33%	23.26%	42.39%	24.52%	32.16%	43.33%
InternLM-XComposer2							
With Annotations	41.23%	20.51%	30.23%	35.87%	20.60%	25.88%	43.33%
Without Annotations	42.54%	38.46%	25.58%	31.52%	23.14%	24.31%	50.00%
CogAgent-VQA							
With Annotations	26.87%	8.97%	0.00%	35.87%	20.32%	23.69%	40.00%
Without Annotations	23.35%	6.41%	2.33%	29.35%	27.71%	29.32%	43.33%
QwenVL							
With Annotations	18.94%	6.41%	2.33%	16.30%	26.61%	28.92%	33.33%
Without Annotations	22.91%	8.97%	2.33%	17.39%	23.36%	25.30%	13.33%

Table 29: China results for all models in the study with EER prompt for continuous maps only

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall.	Rank RWP
GPT-4o							
With Annotations	70.26%	44.68%	51.02%	71.28%	38.64%	45.61%	49.17%
Without Annotations	68.85%	39.38%	53.06%	65.71%	41.58%	47.02%	52.04%
Hatched	49.78%	16.14%	26.83%	26.67%	26.96%	32.60%	60.19%
Gemini 1.5 Flash							
With Annotations	65.03%	54.65%	51.02%	63.10%	38.73%	51.43%	53.80%
Without Annotations	65.70%	59.55%	42.86%	64.48%	33.26%	44.04%	43.61%
Hatched	49.36%	56.20%	51.22%	53.95%	20.51%	35.35%	43.52%
Idefics2							
With Annotations	54.79%	61.11%	18.37%	29.66%	26.64%	42.99%	42.22%
Without Annotations	55.23%	59.09%	24.49%	30.69%	29.61%	43.72%	40.37%
Hatched	52.79%	56.98%	12.20%	25.00%	22.99%	42.09%	45.37%
Intern LM							
With Annotations	52.78%	35.86%	32.65%	20.00%	22.63%	33.07%	42.22%
Without Annotations	52.78%	42.93%	44.90%	22.07%	20.81%	30.12%	38.33%
Hatched	45.49%	31.40%	53.66%	17.11%	25.84%	31.31%	36.11%
CogAgent-VQA							
With Annotations	44.03%	42.23%	24.49%	23.40%	19.79%	33.00%	41.67%
Without Annotations	39.34%	42.23%	22.45%	25.18%	20.16%	27.31%	43.89%
Hatched	34.67%	41.34%	24.39%	20.00%	21.66%	25.09%	41.67%
QwenVL							
With Annotations	37.72%	20.19%	3.19%	6.82%	19.24%	28.59%	33.33%
Without Annotations	35.09%	16.67%	5.32%	8.18%	18.35%	28.48%	35.56%
Hatched	32.42%	9.46%	3.92%	2.46%	18.42%	24.32%	33.33%

Table 30: USA results for all models in the study with zero-shot COT prompt for discrete maps only

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall.	Rank RWP
GPT-4o							
With Annotations	65.57%	62.02%	51.02%	60.64%	37.78%	45.21%	48.33%
Without Annotations	62.30%	59.60%	59.18%	60.99%	39.00%	44.82%	40.93%
Hatched	36.89%	30.71%	21.95%	28.00%	17.48%	24.36%	55.56%
Gemini 1.5 Flash							
With Annotations	66.74%	61.19%	51.02%	62.41%	42.71%	50.28%	50.65%
Without Annotations	64.17%	62.28%	40.82%	61.70%	37.32%	45.05%	44.63%
Hatched	48.44%	40.55%	39.02%	40.00%	18.80%	34.62%	32.87%
Idefics2							
With Annotations	56.57%	27.02%	6.12%	28.97%	23.80%	46.41%	40.00%
Without Annotations	53.45%	26.26%	12.24%	26.21%	21.58%	45.83%	36.67%
Hatched	56.65%	25.19%	14.63%	19.74%	21.24%	44.87%	42.59%
Intern LM							
With Annotations	49.67%	22.22%	16.33%	20.69%	24.41%	35.13%	41.11%
Without Annotations	47.66%	33.08%	26.53%	19.66%	25.74%	36.71%	36.11%
Hatched	47.21%	32.17%	31.71%	11.84%	17.72%	26.60%	45.37%
CogAgent-VQA							
With Annotations	28.10%	20.21%	0.00%	17.38%	19.75%	31.08%	43.33%
Without Annotations	26.70%	30.05%	14.29%	16.67%	20.99%	27.36%	42.04%
Hatched	28.89%	26.77%	9.76%	10.67%	11.15%	15.93%	49.07%
QwenVL							
With Annotations	23.46%	16.67%	6.38%	5.45%	19.88%	32.38%	27.78%
Without Annotations	29.82%	9.62%	3.19%	9.09%	21.28%	30.22%	23.89%
Hatched	29.67%	14.19%	5.88%	3.28%	19.94%	26.18%	20.37%

Table 31: USA results for all models in the study with EER prompt for discrete maps only

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall.	Rank RWP
GPT-4o							
With Annotations	71.52%	40.06%	35.48%	55.75%	49.94%	49.17%	53.70%
Without Annotations	66.45%	40.92%	30.85%	53.54%	46.23%	47.09%	55.56%
Hatched	49.72%	28.38%	31.37%	30.77%	34.19%	36.27%	53.57%
Gemini 1.5 Flash							
With Annotations	56.36%	38.49%	24.47%	40.27%	34.55%	45.11%	38.69%
Without Annotations	58.99%	37.39%	23.40%	35.84%	36.25%	46.21%	45.83%
Hatched	52.75%	48.65%	23.53%	34.38%	38.95%	47.33%	38.89%
Idefics2							
With Annotations	54.17%	43.59%	13.83%	28.76%	32.19%	38.29%	45.24%
Without Annotations	56.58%	43.48%	15.96%	23.45%	28.04%	36.26%	48.81%
Hatched	47.80%	45.95%	17.65%	20.31%	23.31%	30.64%	31.11%
Intern LM							
With Annotations	56.80%	32.37%	17.02%	13.27%	20.14%	24.02%	35.71%
Without Annotations	53.51%	34.08%	14.89%	13.27%	27.84%	35.84%	39.29%
Hatched	51.65%	24.32%	21.57%	24.74%	29.83%	44.44%	44.44%
CogAgent-VQA							
With Annotations	43.27%	25.32%	9.57%	16.81%	19.62%	26.32%	41.36%
Without Annotations	44.37%	29.70%	10.64%	15.49%	22.64%	28.86%	38.89%
Hatched	43.09%	29.50%	7.84%	4.69%	17.30%	20.32%	46.03%
QwenVL							
With Annotations	37.75%	22.33%	4.26%	6.64%	17.00%	23.60%	17.31%
Without Annotations	35.10%	16.67%	5.32%	8.85%	15.50%	19.55%	35.19%
Hatched	32.04%	9.46%	5.88%	2.34%	19.09%	22.55%	21.43%

Table 32: India results for all models in the study with zero-shot COT prompt for discrete maps only

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall.	Rank RWP
GPT-4o							
With Annotations	65.12%	52.46%	40.00%	52.23%	49.88%	51.51%	61.52%
Without Annotations	65.34%	51.82%	40.43%	45.13%	46.98%	47.30%	53.09%
Hatched	43.65%	39.86%	33.33%	28.13%	29.46%	30.76%	61.90%
Gemini 1.5 Flash							
With Annotations	61.37%	37.82%	23.40%	39.38%	29.01%	37.32%	35.60%
Without Annotations	62.69%	42.52%	24.47%	43.81%	39.13%	51.44%	60.26%
Hatched	58.01%	40.32%	33.33%	31.25%	38.01%	53.24%	42.86%
Idefics2							
With Annotations	55.26%	36.54%	8.51%	21.24%	32.65%	48.75%	48.81%
Without Annotations	52.41%	38.25%	11.70%	21.68%	28.17%	46.45%	39.29%
Hatched	54.40%	29.73%	5.88%	12.50%	26.31%	37.28%	26.67%
Intern LM							
With Annotations	51.32%	25.96%	15.96%	13.72%	24.20%	27.80%	53.57%
Without Annotations	52.41%	28.85%	8.51%	13.72%	25.83%	26.68%	47.62%
Hatched	48.35%	13.51%	17.65%	1.56%	17.70%	17.28%	44.44%
CogAgent-VQA							
With Annotations	30.46%	10.90%	4.26%	8.41%	16.55%	19.88%	40.12%
Without Annotations	33.33%	13.46%	4.26%	9.29%	17.52%	19.84%	42.39%
Hatched	31.49%	9.46%	0.00%	6.25%	15.51%	17.84%	29.37%
QwenVL							
With Annotations	23.18%	17.95%	5.32%	5.31%	11.64%	20.43%	29.01%
Without Annotations	30.02%	9.62%	1.06%	8.85%	16.61%	23.10%	32.10%
Hatched	29.83%	14.19%	3.92%	3.13%	17.13%	21.81%	16.67%

Table 33: India results for all models in the study with EER prompt for discrete maps only

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall.	Rank RWP
GPT-4o							
With Annotations	66.97%	47.53%	50.52%	59.40%	53.93%	57.56%	46.58%
Without Annotations	68.10%	41.61%	43.18%	61.27%	44.17%	47.32%	50.48%
Hatched	45.11%	20.56%	27.78%	18.03%	33.33%	34.40%	39.58%
Gemini 1.5 Flash							
With Annotations	62.27%	51.83%	13.40%	52.97%	22.76%	38.96%	53.63%
Without Annotations	60.23%	51.83%	13.40%	53.81%	31.56%	44.84%	41.03%
Hatched	53.80%	55.41%	22.22%	40.32%	29.61%	45.41%	60.42%
Idefics2							
With Annotations	54.09%	50.54%	21.65%	34.32%	21.67%	29.91%	46.15%
Without Annotations	56.82%	48.17%	19.59%	30.93%	21.86%	28.73%	49.36%
Hatched	56.52%	47.62%	16.67%	14.52%	23.01%	33.33%	34.38%
Intern LM							
With Annotations	54.09%	38.39%	19.59%	28.81%	22.98%	28.02%	41.67%
Without Annotations	53.86%	38.49%	22.68%	26.27%	27.28%	33.84%	39.74%
Hatched	50.54%	32.90%	22.22%	4.84%	27.37%	31.62%	47.92%
CogAgent-VQA							
With Annotations	48.64%	28.39%	20.62%	22.03%	21.32%	27.91%	48.08%
Without Annotations	46.47%	28.60%	11.34%	30.34%	29.08%	31.83%	26.92%
Hatched	47.28%	42.64%	11.11%	9.84%	26.39%	32.05%	31.94%
QwenVL							
With Annotations	37.81%	21.18%	3.09%	9.40%	23.92%	27.90%	21.79%
Without Annotations	36.90%	23.66%	2.06%	18.38%	22.56%	27.69%	19.23%
Hatched	40.76%	17.32%	11.11%	3.28%	23.34%	27.35%	12.50%

Table 34: China results for all models in the study with zero-shot COT prompt for discrete maps only

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall.	Rank RWP
GPT-4o							
With Annotations	63.33%	60.65%	45.36%	59.83%	43.47%	46.48%	56.62%
Without Annotations	64.01%	58.39%	49.48%	54.27%	47.24%	50.93%	60.26%
Hatched	44.02%	31.39%	24.07%	21.31%	33.01%	37.82%	38.15%
Gemini 1.5 Flash							
With Annotations	61.50%	54.09%	24.74%	52.14%	23.01%	39.54%	49.54%
Without Annotations	59.23%	53.23%	28.87%	53.42%	29.06%	41.46%	35.90%
Hatched	47.83%	48.92%	25.93%	36.07%	30.84%	50.43%	35.42%
Idefics2							
With Annotations	53.86%	29.25%	18.56%	28.39%	26.63%	39.78%	28.21%
Without Annotations	55.00%	29.78%	19.59%	26.69%	25.70%	37.12%	37.18%
Hatched	52.17%	18.18%	12.96%	9.68%	34.03%	47.44%	39.58%
Intern LM							
With Annotations	42.50%	23.66%	21.65%	24.15%	20.97%	24.74%	41.67%
Without Annotations	45.91%	32.04%	21.65%	19.92%	24.83%	27.86%	46.15%
Hatched	48.37%	20.78%	18.52%	2.42%	30.45%	33.65%	35.42%
CogAgent-VQA							
With Annotations	27.56%	19.35%	5.15%	18.38%	22.25%	25.62%	25.43%
Without Annotations	24.15%	16.77%	9.28%	18.38%	27.61%	30.12%	32.05%
Hatched	33.70%	22.51%	0.00%	8.20%	31.40%	35.58%	14.58%
QwenVL							
With Annotations	21.18%	8.92%	2.06%	8.12%	24.01%	29.35%	24.36%
Without Annotations	24.37%	13.44%	2.06%	8.55%	22.81%	25.98%	16.67%
Hatched	26.09%	7.58%	1.85%	6.56%	20.29%	26.71%	14.58%

Table 35: China results for all models in the study with EER prompt for discrete maps only

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall.
GPT-4o						
With Annotations	78.95%	11.31%	24.00%	0.00%	0.00%	0.00%
Without Annotations	61.05%	7.58%	28.00%	0.00%	0.00%	0.00%
Hactched	50.00%	14.58%	10.53%	0.00%	0.00%	0.00%
Gemini 1.5 Flash						
With Annotations	58.76%	6.67%	24.00%	0.00%	0.00%	0.00%
Without Annotations	69.07%	11.82%	16.00%	0.00%	0.00%	0.00%
Hactched	44.78%	22.92%	15.79%	0.00%	0.00%	0.00%
Idefics2						
With Annotations	44.33%	21.21%	20.00%	0.00%	0.00%	0.00%
Without Annotations	52.58%	30.30%	24.00%	0.00%	0.00%	0.00%
Hactched	38.81%	16.67%	15.79%	0.00%	0.00%	0.00%
InternLM-XComposer2						
With Annotations	40.21%	31.82%	20.00%	0.00%	0.00%	0.00%
Without Annotations	42.27%	45.45%	32.00%	0.00%	0.00%	0.00%
Hactched	26.87%	25.00%	31.58%	0.00%	0.00%	0.00%
CogAgent-VQA						
With Annotations	37.89%	27.27%	8.00%	0.00%	0.00%	0.00%
Without Annotations	36.84%	18.18%	8.00%	0.00%	0.00%	0.00%
Hactched	39.39%	8.33%	21.05%	0.00%	0.00%	0.00%
QwenVL						
With Annotations	41.60%	0.00%	0.00%	0.00%	0.00%	0.00%
Without Annotations	40.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Hactched	35.59%	8.33%	0.00%	0.00%	0.00%	0.00%

Table 36: USA results for all models in the study with zero-shot COT prompt for relative questions only

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall.
GPT-4o						
With Annotations	68.42%	15.76%	24.00%	0.00%	0.00%	0.00%
Without Annotations	52.63%	13.74%	40.00%	0.00%	0.00%	0.00%
Hactched	33.33%	4.17%	0.00%	0.00%	0.00%	0.00%
Gemini 1.5 Flash						
With Annotations	62.11%	32.12%	28.00%	0.00%	0.00%	0.00%
Without Annotations	63.16%	26.36%	12.00%	0.00%	0.00%	0.00%
Hactched	42.42%	16.67%	5.26%	0.00%	0.00%	0.00%
Idefics2						
With Annotations	49.48%	45.45%	4.00%	0.00%	0.00%	0.00%
Without Annotations	45.36%	36.36%	12.00%	0.00%	0.00%	0.00%
Hactched	50.75%	33.33%	10.53%	0.00%	0.00%	0.00%
InternLM-XComposer2						
With Annotations	38.14%	33.33%	8.00%	0.00%	0.00%	0.00%
Without Annotations	42.27%	45.45%	16.00%	0.00%	0.00%	0.00%
Hactched	34.33%	47.92%	10.53%	0.00%	0.00%	0.00%
CogAgent-VQA						
With Annotations	27.37%	48.48%	0.00%	0.00%	0.00%	0.00%
Without Annotations	25.26%	46.97%	4.00%	0.00%	0.00%	0.00%
Hactched	24.24%	20.83%	0.00%	0.00%	0.00%	0.00%
QwenVL						
With Annotations	22.40%	8.33%	3.45%	0.00%	0.00%	0.00%
Without Annotations	29.60%	8.33%	3.45%	0.00%	0.00%	0.00%
Hactched	30.51%	8.33%	14.29%	0.00%	0.00%	0.00%

Table 37: USA results for all models in the study with EER prompt for relative questions only

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall.
GPT-4o						
With Annotations	71.77%	20.83%	35.71%	0.00%	0.00%	0.00%
Without Annotations	68.55%	31.94%	37.93%	0.00%	0.00%	0.00%
Hactched	51.72%	37.50%	21.43%	0.00%	0.00%	0.00%
Gemini 1.5 Flash						
With Annotations	56.80%	18.06%	34.48%	0.00%	0.00%	0.00%
Without Annotations	54.40%	11.11%	37.93%	0.00%	0.00%	0.00%
Hactched	62.71%	25.00%	21.43%	0.00%	0.00%	0.00%
Idefics2						
With Annotations	59.20%	12.50%	10.34%	0.00%	0.00%	0.00%
Without Annotations	62.40%	19.44%	17.24%	0.00%	0.00%	0.00%
Hactched	50.85%	33.33%	35.71%	0.00%	0.00%	0.00%
InternLM-XComposer2						
With Annotations	55.20%	0.00%	13.79%	0.00%	0.00%	0.00%
Without Annotations	53.60%	5.56%	6.90%	0.00%	0.00%	0.00%
Hactched	62.71%	20.83%	14.29%	0.00%	0.00%	0.00%
CogAgent-VQA						
With Annotations	44.35%	12.50%	13.79%	0.00%	0.00%	0.00%
Without Annotations	45.16%	15.28%	10.34%	0.00%	0.00%	0.00%
Hactched	43.10%	11.11%	0.00%	0.00%	0.00%	0.00%
QwenVL						
With Annotations	41.94%	2.78%	3.45%	0.00%	0.00%	0.00%
Without Annotations	39.52%	0.00%	0.00%	0.00%	0.00%	0.00%
Hactched	34.48%	8.33%	7.14%	0.00%	0.00%	0.00%

Table 38: India results for all models in the study with zero-shot COT prompt for relative questions only

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall.
GPT-4o						
With Annotations	68.55%	19.44%	27.59%	0.00%	0.00%	0.00%
Without Annotations	60.48%	31.94%	44.83%	0.00%	0.00%	0.00%
Hactched	43.10%	29.17%	14.29%	0.00%	0.00%	0.00%
Gemini 1.5 Flash						
With Annotations	59.68%	37.50%	13.79%	0.00%	0.00%	0.00%
Without Annotations	66.94%	48.61%	24.14%	0.00%	0.00%	0.00%
Hactched	58.62%	44.44%	28.57%	0.00%	0.00%	0.00%
Idefics2						
With Annotations	52.80%	20.83%	3.45%	0.00%	0.00%	0.00%
Without Annotations	50.40%	34.72%	10.34%	0.00%	0.00%	0.00%
Hactched	59.32%	29.17%	7.14%	0.00%	0.00%	0.00%
InternLM-XComposer2						
With Annotations	44.00%	8.33%	17.24%	0.00%	0.00%	0.00%
Without Annotations	46.40%	8.33%	17.24%	0.00%	0.00%	0.00%
Hactched	35.59%	8.33%	35.71%	0.00%	0.00%	0.00%
CogAgent-VQA						
With Annotations	35.48%	8.33%	6.90%	0.00%	0.00%	0.00%
Without Annotations	37.90%	8.33%	0.00%	0.00%	0.00%	0.00%
Hactched	44.83%	0.00%	0.00%	0.00%	0.00%	0.00%
QwenVL						
With Annotations	22.58%	8.33%	0.00%	0.00%	0.00%	0.00%
Without Annotations	29.84%	8.33%	0.00%	0.00%	0.00%	0.00%
Hactched	31.03%	8.33%	7.14%	0.00%	0.00%	0.00%

Table 39: India results for all models in the study with EER prompt for relative questions only

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall.
GPT-4o						
With Annotations	62.18%	21.43%	56.00%	100.00%	0.00%	0.00%
Without Annotations	63.06%	33.33%	32.00%	0.00%	0.00%	0.00%
Hactched	54.39%	0.00%	10.00%	0.00%	0.00%	0.00%
Gemini 1.5 Flash						
With Annotations	63.87%	28.57%	8.00%	0.00%	0.00%	0.00%
Without Annotations	63.87%	35.71%	16.00%	0.00%	0.00%	0.00%
Hactched	52.63%	23.08%	10.00%	0.00%	0.00%	0.00%
Idefics2						
With Annotations	60.50%	28.57%	28.00%	0.00%	0.00%	0.00%
Without Annotations	59.66%	14.29%	28.00%	0.00%	0.00%	0.00%
Hactched	56.14%	0.00%	30.00%	0.00%	0.00%	0.00%
InternLM-XComposer2						
With Annotations	53.78%	17.86%	4.00%	0.00%	0.00%	0.00%
Without Annotations	54.62%	14.29%	20.00%	0.00%	0.00%	0.00%
Hactched	50.88%	15.38%	10.00%	0.00%	0.00%	0.00%
CogAgent-VQA						
With Annotations	45.38%	21.43%	4.00%	0.00%	0.00%	0.00%
Without Annotations	47.90%	14.29%	4.00%	0.00%	0.00%	0.00%
Hactched	54.39%	26.92%	10.00%	0.00%	0.00%	0.00%
QwenVL						
With Annotations	37.82%	7.14%	0.00%	0.00%	0.00%	0.00%
Without Annotations	41.18%	21.43%	4.00%	0.00%	0.00%	0.00%
Hactched	52.63%	7.69%	10.00%	0.00%	0.00%	0.00%

Table 40: China results for all models in the study with zero-shot COT prompt for relative questions only

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall.
GPT-4o						
With Annotations	57.98%	7.14%	36.00%	100.00%	0.00%	0.00%
Without Annotations	56.30%	21.43%	52.00%	0.00%	0.00%	0.00%
Hactched	50.88%	0.00%	10.00%	0.00%	0.00%	0.00%
Gemini 1.5 Flash						
With Annotations	50.42%	21.43%	12.00%	0.00%	0.00%	0.00%
Without Annotations	51.26%	42.86%	24.00%	0.00%	0.00%	0.00%
Hactched	66.67%	46.15%	0.00%	0.00%	0.00%	0.00%
Idefics2						
With Annotations	57.98%	0.00%	16.00%	0.00%	0.00%	0.00%
Without Annotations	63.03%	14.29%	24.00%	0.00%	0.00%	0.00%
Hactched	49.12%	7.69%	50.00%	0.00%	0.00%	0.00%
InternLM-XComposer2						
With Annotations	43.70%	21.43%	28.00%	0.00%	0.00%	0.00%
Without Annotations	47.90%	21.43%	28.00%	0.00%	0.00%	0.00%
Hactched	54.39%	15.38%	10.00%	0.00%	0.00%	0.00%
CogAgent-VQA						
With Annotations	32.77%	0.00%	4.00%	0.00%	0.00%	0.00%
Without Annotations	38.66%	14.29%	4.00%	0.00%	0.00%	0.00%
Hactched	43.86%	15.38%	0.00%	0.00%	0.00%	0.00%
QwenVL						
With Annotations	21.01%	14.29%	0.00%	0.00%	0.00%	0.00%
Without Annotations	26.89%	0.00%	0.00%	50.00%	0.00%	0.00%
Hactched	24.56%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 41: China results for all models in the study with EER prompt for relative questions only

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall.	Rank RWP
GPT-4o							
With Annotations	67.77%	51.56%	79.17%	71.28%	38.64%	45.61%	49.17%
Without Annotations	71.08%	45.94%	79.17%	65.71%	41.58%	47.02%	52.04%
Hactched	49.69%	16.50%	40.91%	26.67%	26.96%	32.60%	60.19%
Gemini 1.5 Flash							
With Annotations	66.76%	64.24%	79.17%	63.10%	38.73%	51.43%	53.80%
Without Annotations	64.77%	69.09%	70.83%	64.48%	33.26%	44.04%	43.61%
Hactched	51.20%	63.81%	81.82%	53.95%	20.51%	35.35%	43.52%
Idefics2							
With Annotations	57.67%	69.09%	16.67%	29.66%	26.64%	42.99%	42.22%
Without Annotations	55.97%	64.85%	25.00%	30.69%	29.61%	43.72%	40.37%
Hactched	58.43%	66.19%	9.09%	25.00%	22.99%	42.09%	45.37%
InternLM-XComposer2							
With Annotations	56.25%	36.67%	45.83%	20.00%	22.63%	33.07%	42.22%
Without Annotations	55.68%	42.42%	58.33%	22.07%	20.81%	30.12%	38.33%
Hactched	53.01%	32.86%	72.73%	17.11%	25.84%	31.31%	36.11%
CogAgent-VQA							
With Annotations	45.78%	45.31%	41.67%	23.40%	19.79%	33.00%	41.67%
Without Annotations	40.06%	47.19%	37.50%	25.18%	20.16%	27.31%	43.89%
Hactched	32.70%	49.03%	27.27%	20.00%	21.66%	25.09%	41.67%
QwenVL							
With Annotations	36.25%	21.88%	4.62%	6.82%	19.24%	28.59%	33.33%
Without Annotations	33.23%	18.06%	7.69%	8.18%	18.35%	28.48%	35.56%
Hactched	30.89%	9.68%	5.41%	2.46%	18.42%	24.32%	33.33%

Table 42: USA results for all models in the study with zero-shot COT prompt for non-relative questions only

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall.	Rank RWP
GPT-4o							
With Annotations	64.76%	71.56%	79.17%	60.64%	37.78%	45.21%	48.33%
Without Annotations	65.06%	69.06%	79.17%	60.99%	39.00%	44.82%	40.93%
Hactched	38.36%	36.89%	40.91%	28.00%	17.48%	24.36%	55.56%
Gemini 1.5 Flash							
With Annotations	68.07%	67.19%	75.00%	62.41%	42.71%	50.28%	50.65%
Without Annotations	64.46%	69.69%	70.83%	61.70%	37.32%	45.05%	44.63%
Hactched	50.94%	46.12%	68.18%	40.00%	18.80%	34.62%	32.87%
Idefics2							
With Annotations	58.52%	23.33%	8.33%	28.97%	23.80%	46.41%	40.00%
Without Annotations	55.68%	24.24%	12.50%	26.21%	21.58%	45.83%	36.67%
Hactched	59.04%	23.33%	18.18%	19.74%	21.24%	44.87%	42.59%
InternLM-XComposer2							
With Annotations	52.84%	20.00%	25.00%	20.69%	24.41%	35.13%	41.11%
Without Annotations	49.15%	30.61%	37.50%	19.66%	25.74%	36.71%	36.11%
Hactched	52.41%	28.57%	50.00%	11.84%	17.72%	26.60%	45.37%
CogAgent-VQA							
With Annotations	28.31%	14.37%	0.00%	17.38%	19.75%	31.08%	43.33%
Without Annotations	27.11%	26.56%	25.00%	16.67%	20.99%	27.36%	42.04%
Hactched	30.82%	28.16%	18.18%	10.67%	11.15%	15.93%	49.07%
QwenVL							
With Annotations	23.87%	17.36%	7.69%	5.45%	19.88%	32.38%	27.78%
Without Annotations	29.91%	9.72%	3.08%	9.09%	21.28%	30.22%	23.89%
Hactched	29.27%	15.32%	2.70%	3.28%	19.94%	26.18%	20.37%

Table 43: USA results for all models in the study with EER prompt for non-relative questions only

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall.	Rank RWP
GPT-4o							
With Annotations	71.43%	41.67%	35.38%	55.75%	49.94%	49.17%	53.70%
Without Annotations	65.65%	41.67%	27.69%	53.54%	46.23%	47.09%	55.56%
Hactched	48.78%	26.61%	35.14%	30.77%	34.19%	36.27%	53.57%
Gemini 1.5 Flash							
With Annotations	56.19%	40.21%	18.46%	40.27%	34.04%	45.99%	39.29%
Without Annotations	60.73%	39.58%	16.92%	35.84%	35.70%	46.64%	48.21%
Hactched	47.97%	53.23%	27.03%	34.38%	37.79%	46.00%	38.89%
Idefics2							
With Annotations	52.27%	46.18%	15.38%	28.76%	32.19%	38.29%	43.45%
Without Annotations	54.38%	45.49%	15.38%	23.45%	28.04%	36.26%	50.00%
Hactched	46.34%	48.39%	10.81%	20.31%	23.31%	30.64%	31.11%
InternLM-XComposer2							
With Annotations	57.40%	35.07%	18.46%	13.27%	19.26%	23.26%	39.29%
Without Annotations	53.47%	36.46%	16.92%	13.27%	26.96%	36.15%	40.48%
Hactched	46.34%	25.00%	29.73%	0.00%	24.44%	29.14%	46.67%
CogAgent-VQA							
With Annotations	42.86%	26.39%	7.69%	16.81%	19.62%	26.32%	41.36%
Without Annotations	44.07%	30.90%	10.77%	15.49%	22.64%	28.86%	38.89%
Hactched	43.09%	33.06%	10.81%	4.69%	17.30%	20.32%	46.03%
QwenVL							
With Annotations	36.17%	23.96%	4.62%	6.64%	17.00%	23.60%	17.31%
Without Annotations	33.43%	18.06%	7.69%	8.85%	15.50%	19.55%	35.19%
Hactched	30.89%	9.68%	5.41%	2.34%	19.09%	22.55%	21.43%

Table 44: India results for all models in the study with zero-shot COT prompt for non-relative questions only

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall.	Rank RWP
GPT-4o							
With Annotations	63.83%	55.21%	45.45%	52.23%	49.88%	51.51%	61.52%
Without Annotations	67.17%	53.47%	38.46%	45.13%	46.98%	47.30%	53.09%
Hactched	43.90%	41.94%	40.54%	28.13%	29.46%	30.76%	61.90%
Gemini 1.5 Flash							
With Annotations	62.01%	37.85%	27.69%	39.38%	29.01%	37.32%	35.60%
Without Annotations	61.09%	42.01%	24.62%	43.81%	39.13%	51.44%	60.26%
Hactched	57.72%	39.52%	35.14%	31.25%	38.01%	53.24%	42.86%
Idefics2							
With Annotations	56.19%	37.85%	10.77%	21.24%	32.65%	48.75%	48.81%
Without Annotations	53.17%	38.54%	12.31%	21.68%	28.17%	46.45%	39.29%
Hactched	52.03%	29.84%	5.41%	12.50%	26.31%	37.28%	26.67%
InternLM-XComposer2							
With Annotations	54.08%	27.43%	18.46%	13.72%	23.85%	27.66%	55.95%
Without Annotations	54.68%	30.56%	10.77%	13.72%	25.64%	28.30%	47.02%
Hactched	54.47%	14.52%	18.92%	1.56%	17.70%	17.77%	44.44%
CogAgent-VQA							
With Annotations	28.57%	11.11%	3.08%	8.41%	16.55%	19.88%	40.12%
Without Annotations	31.61%	13.89%	6.15%	9.29%	17.52%	19.84%	42.39%
Hactched	25.20%	11.29%	0.00%	6.25%	15.51%	17.84%	29.37%
QwenVL							
With Annotations	23.40%	18.75%	7.69%	5.31%	11.64%	20.43%	29.01%
Without Annotations	30.09%	9.72%	1.54%	8.85%	16.61%	23.10%	32.10%
Hactched	29.27%	15.32%	2.70%	3.13%	17.13%	21.81%	16.67%

Table 45: India results for all models in the study with EER prompt for non-relative questions only

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall.	Rank RWP
GPT-4o							
With Annotations	68.75%	50.12%	48.61%	59.05%	53.93%	57.56%	46.58%
Without Annotations	70.07%	42.35%	47.62%	61.88%	44.17%	47.32%	50.48%
Hactched	40.94%	24.74%	31.82%	18.03%	33.33%	34.40%	39.58%
Gemini 1.5 Flash							
With Annotations	61.68%	54.14%	15.28%	53.42%	22.76%	38.96%	53.63%
Without Annotations	58.88%	53.43%	12.50%	54.27%	31.56%	44.84%	41.03%
Hactched	54.33%	61.98%	25.00%	40.32%	29.61%	45.41%	60.42%
Idefics2							
With Annotations	51.71%	52.72%	19.44%	34.62%	21.67%	29.91%	46.15%
Without Annotations	55.76%	51.54%	16.67%	31.20%	21.86%	28.73%	49.36%
Hactched	56.69%	57.29%	13.64%	14.52%	23.01%	33.33%	34.38%
InternLM-XComposer2							
With Annotations	54.21%	40.43%	25.00%	29.06%	22.98%	28.02%	41.67%
Without Annotations	53.58%	40.90%	23.61%	26.50%	27.28%	33.84%	39.74%
Hactched	50.39%	36.46%	25.00%	4.84%	27.37%	31.62%	47.92%
CogAgent-VQA							
With Annotations	48.64%	28.39%	20.62%	22.03%	21.32%	27.91%	48.08%
Without Annotations	45.94%	30.02%	13.89%	30.60%	29.08%	31.83%	26.92%
Hactched	44.09%	45.83%	11.36%	9.84%	26.39%	32.05%	31.94%
QwenVL							
With Annotations	37.81%	22.58%	4.17%	9.48%	23.92%	27.90%	21.79%
Without Annotations	35.31%	23.88%	1.39%	18.53%	22.56%	27.69%	19.23%
Hactched	35.43%	19.27%	11.36%	3.28%	23.34%	27.35%	16.67%

Table 46: China results for all models in the study with zero-shot COT prompt for non-relative questions only

Map Type	Binary Accuracy	Single Recall	Count Accuracy	Range Accuracy	List Precision	List Recall.	Rank RWP
GPT-4o							
With Annotations	65.31%	65.96%	48.61%	59.48%	43.47%	46.48%	56.62%
Without Annotations	66.88%	62.06%	48.61%	54.74%	47.24%	50.93%	60.26%
Hactched	40.94%	37.76%	27.27%	21.31%	33.01%	37.82%	38.15%
Gemini 1.5 Flash							
With Annotations	65.63%	57.33%	29.17%	52.59%	23.01%	39.54%	49.54%
Without Annotations	62.19%	54.26%	30.56%	53.88%	29.06%	41.46%	35.90%
Hactched	39.37%	49.48%	31.82%	36.07%	30.84%	50.43%	35.42%
Idefics2							
With Annotations	52.34%	32.15%	19.44%	28.63%	26.63%	39.78%	28.21%
Without Annotations	52.02%	31.32%	18.06%	26.92%	25.70%	37.12%	37.18%
Hactched	53.54%	20.31%	4.55%	9.68%	34.03%	47.44%	39.58%
InternLM-XComposer2							
With Annotations	42.06%	23.88%	19.44%	24.36%	20.97%	24.74%	41.67%
Without Annotations	45.17%	33.10%	19.44%	20.09%	24.83%	27.86%	46.15%
Hactched	45.67%	21.88%	20.45%	2.42%	30.45%	33.65%	35.42%
CogAgent-VQA							
With Annotations	25.62%	21.28%	5.56%	18.53%	22.25%	25.62%	25.43%
Without Annotations	18.75%	17.02%	11.11%	18.53%	27.61%	30.12%	32.05%
Hactched	29.13%	23.96%	0.00%	8.20%	31.40%	35.58%	14.58%
QwenVL							
With Annotations	21.25%	8.39%	2.78%	8.19%	24.01%	29.35%	24.36%
Without Annotations	23.44%	14.78%	2.78%	8.19%	22.81%	25.98%	19.23%
Hactched	26.77%	9.11%	2.27%	6.56%	20.29%	26.71%	14.58%

Table 47: China results for all models in the study with EER prompt for non-relative questions only