MAPQA: A MAP-QUESTION-ANSWERING BENCHMARK FOR VISUAL LANGUAGE MODEL REASONING

Christian M. A. Arnold, Andrew Alini, Jonathan Wang, Pieter M. Feenstra & Conner Arnold
Computer Science Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
cmarnold@csail.mit.edu, andrew.alini@ll.mit.edu,
{jwang27, maxfeen, carnold}@csail.mit.edu

Jan DeWitt, Natalie C. Ritsema, Jung Hyun Yae, Boris Katz, Andrei Barbu, & Brian Cheung
Computer Science Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{jkdewitt, nritsema}@csail.mit.edu, jyae@mit.edu
{boris, abarbu}@csail.mit.edu, cheungb@mit.edu

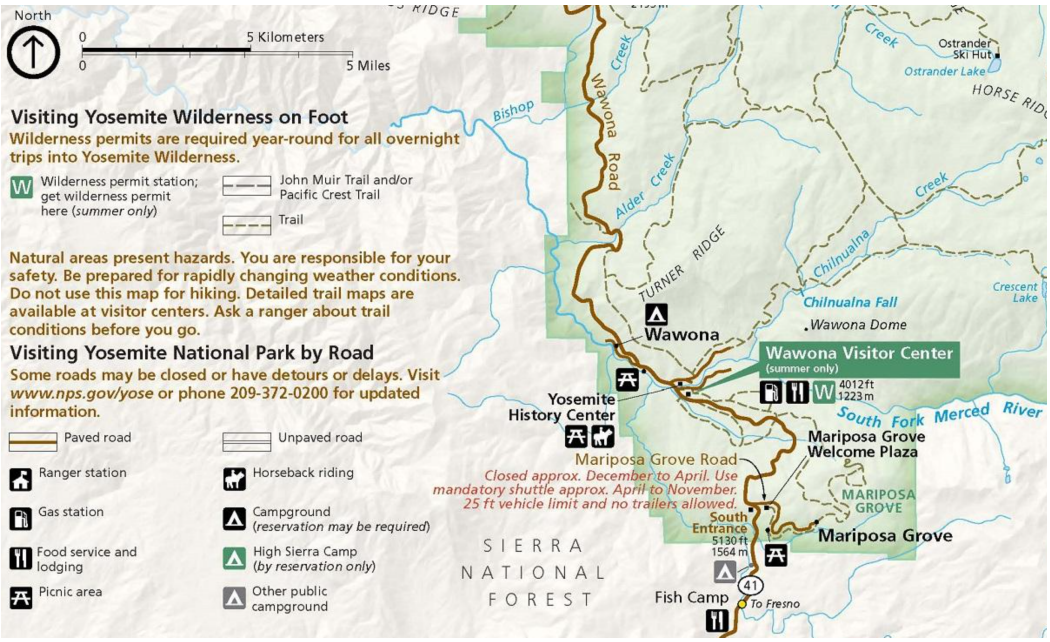
ABSTRACT

Maps are central to how humans make sense of the world, from navigation and environmental monitoring to military planning and historical interpretation. Yet despite rapid progress in large multimodal models (LMMs), these systems continue to struggle with interpreting maps–an essential skill for visual reasoning that goes beyond pattern recognition and text extraction. To further progress toward this capability, we introduce MapQA, the first large-scale benchmark specifically designed to evaluate LMMs on map understanding. MapQA contains over 4,200 carefully curated, open-ended question–answer pairs spanning diverse map types, each constructed to require reasoning directly from the map rather than relying on memorized world knowledge. Benchmark questions were generated through a scalable human-in-the-loop process to ensure quality, and evaluated using an LLM-as-a-judge protocol aligned with human judgments. Our experiments show that while humans answer over 91% of questions correctly, state-of-the-art proprietary models achieve barely half that performance, with open-source models typically below 30%. These findings highlight a substantial gap between human and machine map understanding, underscoring the need for benchmarks like MapQA to guide future progress in multimodal reasoning.

1 INTRODUCTION

LMMs have shown remarkable improvement in open-domain visual question answering across a broad range of tasks, such as math, photography, and interpreting charts and diagrams (Agrawal et al., 2024; Abdin et al., 2024; OpenAI, 2024). A wide variety of benchmark datasets, spanning from general perception (Yue et al., 2024; Ying et al., 2024) to specialized tasks (Mathew et al., 2021; Lu et al., 2023; Masry et al., 2022), exist to evaluate LMMs' capabilities. However, to achieve strong perception across domains, LMMs must extend into more specialized and challenging areas that often lack comprehensive benchmark datasets.

Map understanding is one such domain of particular importance. This is because maps aid our understanding of the world and shape our perception through the narratives they construct (Marthew, 2023). They have broad applications throughout society, including ecological and geoscientific maps for understanding Earth systems (Teixeira & Teodoro, 2021), tactical cartography for military situational awareness (Liebenberg et al., 2016), and street maps for navigation (Nakhimovsky et al., 2010). Maps are widely used because they act as external cognitive aids, offloading abstract and



Context: The following is a zoomed in and cropped version of a map of Yosemite National Park from the Library of Congress. The original image is 3000 by 3175 pixels and we have zoomed in on the particular region of interest to the following question.
Question: What symbols are indicated as available amenities at Yosemite History Center?
Correct Answer: Picnic areas and horseback riding.
GPT-5 Answer with the map: Information and wheelchair-accessible symbols.
GPT-5 Answer without the map: Ranger station and Information center.

Figure 1: Map of Yosemite National Park: Image with context, question, ground truth, and GPT-5 incorrect answers. Despite the legend for the symbols being next to the label for Yosemite History Center, every model tested completely ignores the legend and hallucinates an answer.

complex concepts, as well as spatial relationships, into symbolic representations that humans can quickly interpret (Casati, 2023).

Despite recent advances, LMMs continue to struggle with understanding and interpreting maps. Recent work (Tong et al., 2024) sheds light on the sources of this difficulty by identifying nine common categories of failure in visual question answering: (i) orientation and direction, (ii) presence of specific features, (iii) state and condition, (iv) quantity and count, (v) positional and relational context, (vi) color and appearance, (vii) structural and physical characteristics, (viii) text, and (ix) viewpoint and perspective. All of these categories are foundational to map understanding, as they correspond to the core visual encodings and spatial relationships through which maps convey meaning. In addition, prior work (Goyal et al., 2017; Rohrbach et al., 2018) shows that LMMs frequently underutilize visual evidence during reasoning, allowing language inputs and learned world knowledge to dominate their predictions. When applied to maps, where critical information is encoded symbolically and relationally, this tendency can cause models to favor linguistically plausible or historically accurate answers that are inconsistent with the visual content, resulting in errors or hallucinations. Figure 1 illustrates this failure mode, in which GPT-5 produces an answer driven by historical knowledge rather than by evidence present in the map.

To address these issues, we present the MapQA Benchmark, a new vision-based question answering dataset designed to evaluate LMMs' reasoning skills on a diverse set of complex, information-rich maps. The questions in the dataset were specifically designed and curated to be answerable only with the map, in order to directly test map understanding and prevent LMMs from taking shortcuts using intrinsic knowledge from language. The dataset was created through human-in-the-loop LLM generation and careful curation of generated outputs and human-annotated answers. We evaluate the

benchmark on a wide variety of state-of-the-art open-source and proprietary models and demonstrate their limitations in map question answering.

2 Related Work

Map Benchmark Datasets With the growth of research into LMMs, single-task (Antol et al., 2015; Goyal et al., 2017; Marino et al., 2019; Lin et al., 2014) and general-perception (Yin et al., 2023; Yu et al., 2023; Liu et al., 2024b; Yue et al., 2024) multimodal datasets have become increasingly important for training and evaluation. While the latter provide more holistic coverage, single-task evaluations remain critical for probing isolated skills of LMMs. This is because domain-specific sub-capabilities may be overlooked if certain question types are not framed appropriately or are omitted entirely (Li et al., 2025). Within this space, map question-answering remains limited in scope. Extensive map datasets exist across domains such as geography (U.S. Geological Survey, 2025b; Kelso & Patterson, 2009; Fick & Hijmans, 2017), remote sensing (U.S. Geological Survey, 2025a; Hansen et al., 2013; Van Etten et al., 2018), transportation (OpenStreetMap contributors, 2017; Chang et al., 2019), and history (Library of Congress, 2025; David Rumsey Map Collection, 2025; Kramm et al., 2025). However, these are largely uncurated collections of raw data without labels or paired QA annotations needed to evaluate LMMs. Prior work has examined visual question answering for maps, but has typically restricted its focus to remote sensing, geometric questions, choropleth maps, and transportation maps rather than broader applications (Lobry et al., 2020; Chen et al., 2021a; Chang et al., 2022; Li et al., 2025; Punjani et al., 2018; Kefalidis et al., 2023).

Algorithmic QA-Judge Building algorithms that can effectively judge answers is an active area of research. Two common techniques are key-phrase matching and using LLMs as judges. Key-phrase matching aligns key phrases, words, or tokens in the ground-truth answer with those in the generated response. Early work primarily relied on regular expressions (Yue et al., 2024; Sychev, 2023), while later approaches assign token importance through weighting (Lee et al., 2021) or neural classifiers (Bulian et al., 2022). These methods still face challenges such as varying answer granularity, multiple candidate answers, and lexical mismatches. Some approaches attempt to address these issues by expanding ground-truth knowledge sets (Si et al., 2021) or by applying natural language inference models to reformulated questions and answers (Chen et al., 2021b; Honovich et al., 2021). However, they remain limited when key phrases fall outside the dataset.

LLMs as judges have seen growing interest in recent years for their potential to overcome the flaws of key-phrase matching. Popular approaches involve directly prompting an LLM to evaluate QA responses (Kamalloo et al., 2023; Myrzakhan et al., 2024). For example, Myrzakhan et al. (2024) propose prompts for evaluating both open-set and multiple-choice answers, which we adopt for our MapQA evaluation task. Other work has examined LLM-based evaluation of natural language generation and summarization (Liu et al., 2023; Sottana et al., 2023), while Zheng et al. (2023) explore using an LLM to rank chatbot answers by correctness and quality. Since large models are expensive to run, recent efforts have sought to reduce costs by finetuning smaller LLMs on judgments from larger models such as GPT-4o (Zhu et al., 2023).

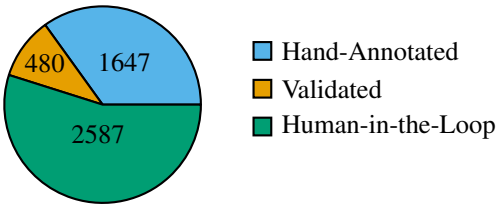
3 The MapQA Benchmark

3.1 Overview

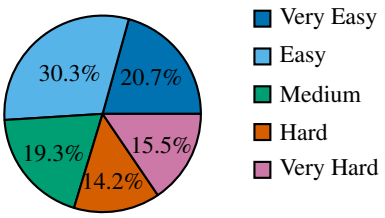
We introduce the MapQA Benchmark, a novel vision-based question-answer benchmark to test LMMs ability to reason over information-dense images, and evaluate current model capabilities. Human perception relies on vision at its core, and functions as a primary method of communication. Where text can be laborous to read and interpret, a map is often easy to understand and presents massive quantities of information to a user in a manner that is quickly interpretable while also not being overwhelming. To further improve models and thoroughly test their capabilities, a rigorous vision-based dataset is needed to compare models against human commonsense and reasoning. The MapQA dataset encompasses four primary map types—Military, Natural World, Urban, and Aviation—spanning 24 subcategories in total. See section B for more details. Each type poses distinct challenges for visual understanding, making them valuable for systematic evaluation.

Table 1: Key statistics of the MapQA benchmark.

Statistics	**Number**
Total Questions	4,234
Total Map Types/Subcategories	4/24
Open-Answer Questions	4,234 (100.0%)
Questions with Multiple Images	811 (19.2%)
Average question word length	12.7
Average answer word length	4.1
Total Unique Images	578
Largest image size (pixels)	6000 × 6000
Smallest image size (pixels)	546 × 393
Average image size (pixels)	4782 × 4198
Standard deviation (pixels)	1545 × 1289



(a) Question split distribution (counts).



(b) Difficulty distribution (percent).

Figure 2: Overview of MapQA benchmark statistics.

MapQA is composed of 4,234 questions, with each one designed to be unanswerable without the aid of a map. This ensures we eliminate bias from parameterized knowledge about any given region. For example, questions like "Which airport is closest to Chicago O'Hare?" are easily answerable without a map because the knowledge is very likely to already be in a given model's training data. We also minimize the number of questions that can be answered strictly via Optical Character Recognition (OCR), as these questions are simplying reading text on a given map rather than reasoning over the map and it's implications. Lastly, the dataset must be answerable by humans to a reasonable degree of accuracy. In our case, we found that among questions that were tested, humans answer 91.4% of them correctly.

The MapQA dataset uses an open-answer format to rigorously evaluate a model's reasoning ability rather than its capacity to select the most plausible option from a fixed set of multiple-choice answers. However, open-answer evaluation introduces additional challenges, particularly in determining whether a model's response is semantically correct. To address this, we employ an LLM-as-a-Judge framework, which has been shown to provide more accurate semantic similarity judgments than traditional algorithmic evaluation methods (Zhu et al., 2023). When compared to other models, we found GPT4o to be the most accurate and cost effective. After calibrating GPT4o to over two thousand human answers to MapQA questions, we fit GPT4o judgements to a linear model $y = \alpha + \beta x$ and found $\alpha = 0.035$ and $\beta = 0.921$, indicating that GPT4o when used as LLM-As-A-Judge, while not perfect, is a very strong indicator of ground-truth accuracy.

To score model predictions we adapt the LLM-As-A-Judge framework modifying the system so that it issues a binary correct/incorrect verdict, tolerating minor phrasing differences as well as numerical slack ($\pm 10\%$) when judging measurement-heavy questions. These instructions diverge from the original paper's open form rationales by tightly constraining the response and explicitly allowing the judge to reconcile student and reference answers that describe the same cartographic features in different words.

3.2 Curation Process

Hand-Curation The MapQA benchmark is carefully curated with the previously stated inclusion criteria in mind. For the first 2,000 questions, our co-authors hand-curated every questions and cross-scrutinized each for clarity and interpretability before culling any question that is answerable without a map. Culling was repeated until accuracy fell below 3.5%–the previously computed α for GPT4o when used in LLM-As-A-Judge. After scrutinizing and culling, 1.6k questions remained over 172 unique images of maps.

Human-in-the-loop Curation Manual curation is inherently labor-intensive and time-consuming, and it can be difficult to monitor quality over a large dataset when annotators tend to converge to-

wards questions that follow repetitive and formulaic patterns. While crowdsourcing offers a potential alternative, our attempts revealed significant limitations: crowdworkers lacked familiarity with the capabilities and shortcomings of modern LMMs, and frequently generated questions that were either answerable without the given map or depended primarily on optical character recognition (OCR). Consequently, fewer than 5% of crowdworker-generated questions satisfied our inclusion criteria, underscoring the need for an alternative strategy.

Instead, we implemented an automated human-in-the-loop curation strategy that leverages few-shot examples for each map type. Starting with an image of a map, we use GPTo4-mini to generate 30 candidate questions based on 50 example questions from our hand-curated set. Those questions are presented to two annotators who have previously been evaluated for their expertise on the hand-curated set. LLM-As-As-Judge then evaluates whether the answers provided by each annotator is similar. If the answers are similar, then the question-answer pairs are culled until LLM-As-Judge Accuracy without images falls below 3.5%. Once this is complete, the questions-answer-map triplets are accepted into the benchmark set. If the two responses are not found to be similar, they advance to a manual review from our research team where, after manual review and possible revision, the question-answer may be ultimately accepted or rejected. The details of this process are shown in fig. 3. Using this automated curation strategy, we added an additional 2.6k questions for a total of 4.2k questions. The annotation instructions can be found in Appendix Section H.

Finding Expert Humans We determine an individual annotator's aptitude by first surveying their response to a random subset of our original hand-curated set from the research team. For each map type, we randomly sample approximately 120 questions to determine the calibrated accuracy, p, for that annotator. Those with a $p > 0.95$ are considered 'experts' in that category.

We calculate the probability given that two experts independently arrive at the same answer below:

$$\text{For two experts, each with accuracy } p, \quad P(\text{correct} \mid \text{2 unanimous}) = \frac{p^2}{p^2+(1-p)^2}.$$

$$P(\text{correct} \mid p = 0.90) = \frac{0.90^2}{0.90^2 + 0.10^2} \approx \frac{0.81}{0.81 + 0.01} \approx 0.9878,$$

Thus, if responses are agreed on by our annotators that average a 91.4% accuracy on the 120 question set, we have strong statitical confidence that the concurred answer to automatically generated questions are correct.

Determining Difficulty During the evaluation of annotators on our hand-curated set, we also measure the length of time required to respond to each question. This data is sorted into quintiles from which we derive five difficulty categories: Very Easy, Easy, Medium, Hard, and Very Hard. This relationship is then used to determine the difficulty of all subsequent questions created via the human-in-the-loop data curation process.

4 EXPERIMENTS

We consider various large multimodal models to evaluate the MapQA benchmark, with a near even split between open-source models and proprietary models. We choose the latest model version for each family for our evaluation, and use a variety of model parameter sizes within each family to

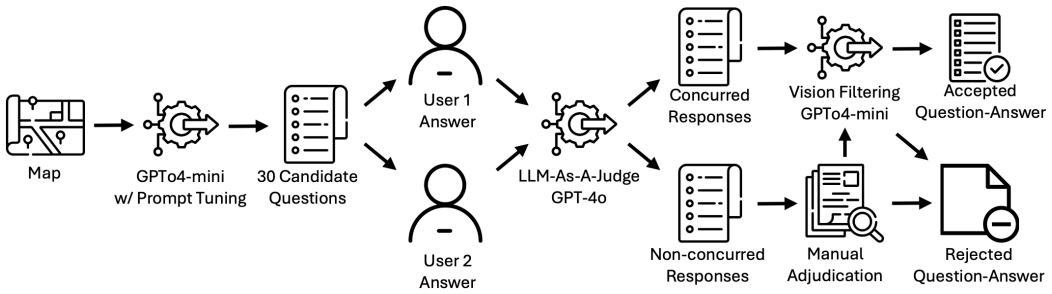


Figure 3: Automated Human-in-the-loop Curation Flow Diagram.

observe how performance scales with model size. We perform zero-shot evaluation with a system prompt to evaluate the model capability to generate correct answers to the benchmark questions without finetuning. The system prompt is shown in fig. 4. For open-source models, experiments are conducted with two different types of GPUs depending on model size. For very large models, the model is sharded across 4 NVIDIA H100 80GB GPUs. When a smaller model is possible, the model is either run on a single NVIDIA A100 40GB GPU or it may be sharded across 4 A100 40GB GPUs.

```
Images presented are tiles from one single image.  The tiles
are presented from top to bottom, left to right.  Tiles are
padded with black pixels.  Answer the question considering
all image tiles.  Your answers should be relatively concise.
```

Figure 4: MapQA System Prompt to answer questions

4.1 Baselines

Open-Source LMMs We evaluate various open-source models of different parameter sizes. (i.) Qwen2.5-VL-3B extends the Qwen-VL series by introducing dynamic resolution processing to handle images of various sizes, enabling perception of spatial scales without traditional normalization techniques (Bai et al., 2025). (ii.) Pixtral-12B has a custom-trained vision encoder that processes images of varying resolutions and aspect ratios (Agrawal et al., 2024).

(iii.) Phi-3.5 Vision Instruct ONNX interweaves vision tokens from CLIP ViT-L/14 and text tokens before passing them into the phi-3.5-mini decoder (Abdin et al., 2024). LLaVA-Next passes visual encodings into an LLM using both fine-grained patch embeddings and a global embedding from the resized image. We evaluate LLaVA-Next with various LLM decoders and sizes: Qwen (iv. LLaVA-Next-34B, v. LLaVA-Next-72B), Vicuna (vi. LLaVA-Next-Vicuna-7B, vii. LLaVA-Next-Vicuna-13B), and Mistral (viii. LLaVA-Next-Mistral-7B) (Liu et al., 2024a). (ix.) Llama-3-11B-VL adds image input support by adding cross-attention adapter weights to a pretrained image model before passing them into a pretrained language model (AI, 2024).

The large images found in MapQA require orders of magnitude more token input than regular text. As a result, some models that can run on the listed devices with text input alone are unable to fit a single sample in the remaining memory of an 80 GB GPU after loading just one-quarter of the model. For this reason, we are unable to benchmark models such as Pixtral-124B, LLaVA-Next-110B, Llama-3-90B, Qwen2.5-VL-32B, and Qwen2.5-VL-72B.

Proprietary LMMs We also consider several proprietary models. (x.) GPT-4o is OpenAI's first omni model, capable of accepting and generating any combination of text, audio, image, and video input (OpenAI, 2024).

(xi.) GPT-4.1 improves on GPT-4o with a longer context window and improved instruction following (OpenAI, 2025a). (xii.) GPT-o4-mini is a compact and faster version of GPT-o4 (OpenAI). (xiii.) GPT-5 is a unified system with a deep reasoning model for harder problems and a router to select the appropriate model given the conversation and prompt (OpenAI, 2025b). (xiv.) GPT-5-mini and (xv.) GPT-5-nano are smaller, more compact versions of GPT-5 (OpenAI, 2025b;c). Anthropic's Claude-Opus-4 is a strong coding model focused on complex reasoning and long-running tasks, while Claude-Sonnet-4 is tuned for improved instruction following in coding and reasoning tasks (Anthropic, 2025). Claude-3.5-Haiku provides lower latency for inference while offering comparable performance to Claude-Opus-3 (Anthropic, 2024).

Humans To assess annotation quality, we evaluated all human annotators on a random sample of 120 hand-curated questions from each map type, yielding a validation set of 600 questions in total. The resulting accuracies, disaggregated by category, are reported in table 2.

Evaluation We report micro-averaged accuracy, treating each question as a single instance and averaging correctness over all items in the benchmark. Because MapQA is open-answer, we assess correctness using an LLM-as-a-Judge protocol (GPT-4o) with the model's intermediate reasoning masked to prevent leakage. Judge outputs are calibrated to human judgments via a linear mapping $y = \alpha + \beta x$ fit on more than two thousand human-rated responses for each map type; the resulting

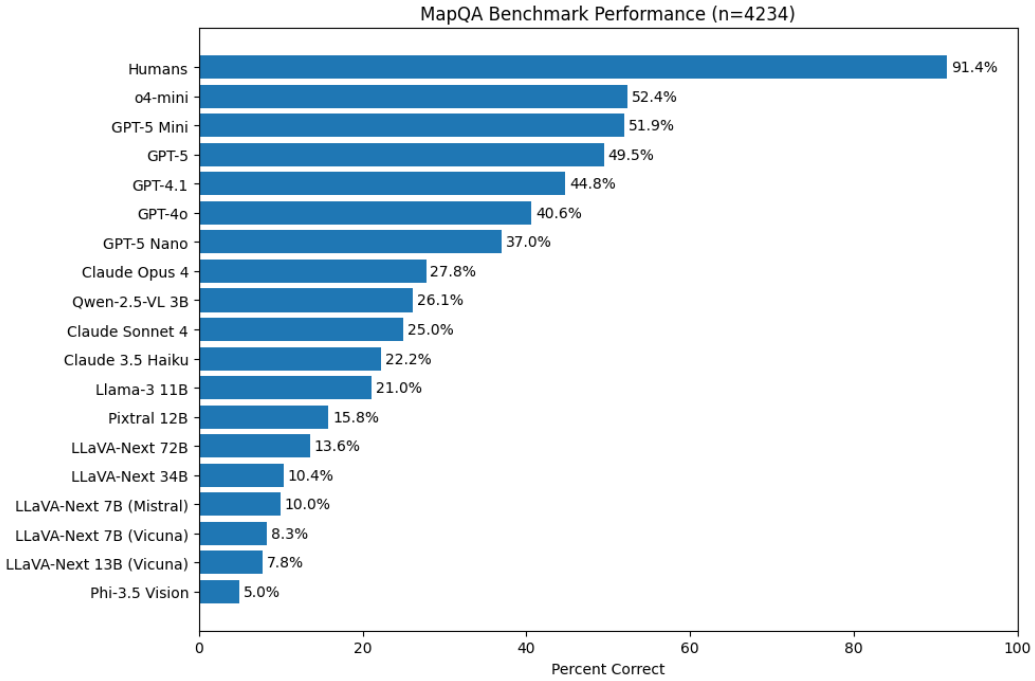


Figure 5: MapQA Benchmark evaluation across all models. Humans significantly outperform the best LMM models.

coefficients are used to inverse-map raw judge scores to calibrated accuracies. Reported accuracies reflect this calibrated, micro-averaged measure over the full evaluation set.

4.2 Benchmark Results

The MapQA benchmark is tested against 18 models and plotted in fig. 5. On aggregate, humans strongly outperform all models with 91.4% accuracy based on performance on the Validated set of questions. The best models against the benchmark are the proprietary models. GPTo4-mini tops the benchmark with 52.5% accuracy. The best performing open-source model is Qwen2-2B with 26.1% outperforming Claude Sonnet 4 and Haiku 3.5, while slightly underperforming Opus 4. There appears to be little difference between OpenAI generations when comparing GPTo4-mini to GPT5-mini, with GPTo4-mini slightly outperforming the latter by 0.6%, and a small improvement from GPT4.1 to GPT5, with a difference of 4.8% separating the two.

Among the Claude family of models, there also does not appear to be large differences between model performance. Based on model pricing, one would assume that Claude Opus 4 would significantly outperform Sonnet 4 and Haiku 3.5. According to the pricing at the time of writing, Opus 4 is 5x more expensive than Sonnet 4, and 18.75x more expensive than Haiku 3.5, however, the difference between Opus 4 and Haiku 3.5 is a meager 5.6%.

Category	Human Accuracy
Military	93.7%
Natural World	90.4%
Urban	90.8%
Aviation	90.9%
Benchmark Average	**91.4%**

Table 2: Average human accuracy across MapQA categories.

5 ANALYSIS

We analyze the benchmark to provide insights that could explain model performance and how they can be modified to further improve accuracy. We explore model size, tokens used, and performance group by difficulty.

5.1 MODEL SIZE INSIGNIFICANCE

We analyze how MapQA performance scales with parameter size of each model. Unfortunately, there is little information available regarding the size of proprietary models, so only comparisons of open-source model accuracy is performed. This does mean that the comparison is limited because no open-source model exceeds 30% accuracy, however, as shown in fig. 6, there is no statistically significant trend. Qwen2.5-VL 2B is the best performing open-source model, but also the smallest, while LlaVA-Next 72B is the largest and shows middling performance.

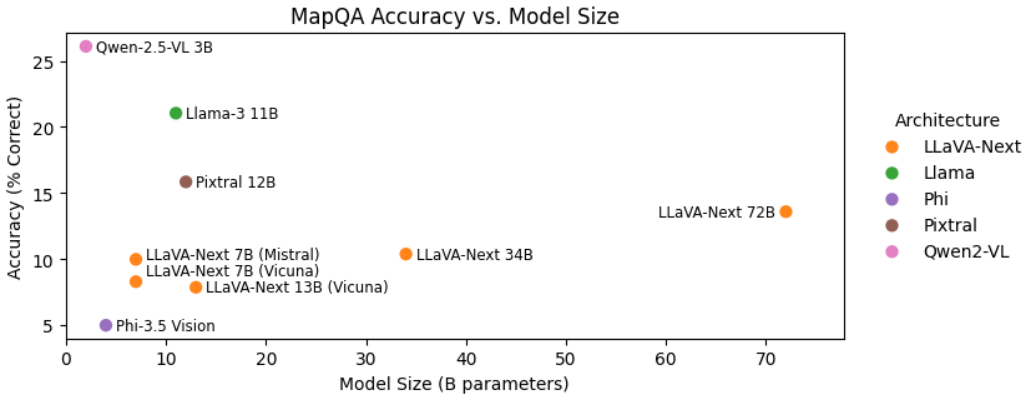


Figure 6: MapQA Benchmark performance compared to model size. Averages are computed per model with each dot representing one model.

5.2 POSITIVE CORELATION WITH TOKEN USAGE

When comparing the average number of tokens used against a model's average accuracy, an interesting trend emerges: As the number of average tokens increases, the accuracy against MapQA also increases with a strong (r=0.6976) correlation that is statistically significant across 4.2k samples, as shown in fig. 7.

Average tokens used are the aggregate of input tokens, chain-of-thought (if any), and output tokens. Unfortunately, the number of chain-of-thought tokens is not individually reported by proprietary models, and is instead categorized like an output token, so deeper analysis is unavailable. In general, models that employ larger image encoder vocabularies and/or use of chain-of-thought achieve higher performance against the MapQA benchmark.

6 CONCLUSION

MapQA establishes the first large-scale, systematically curated benchmark dedicated to evaluating map understanding in large multimodal models. By constructing open-ended, map-dependent questions across diverse domains, MapQA goes beyond text extraction and memorized world knowledge to test genuine spatial reasoning, relational inference, and visual interpretation. Our experiments reveal a striking gap: even the strongest proprietary models fall nearly 40% short of human performance, with open-source models lagging further behind. This persistent shortfall underscores that current LMM progress in charts, diagrams, and general VQA does not yet extend to the unique challenges of cartographic reasoning.

By making these limitations visible, MapQA provides not only a diagnostic tool but also a research agenda. It points to the need for new architectures, training regimes, and evaluation strategies that

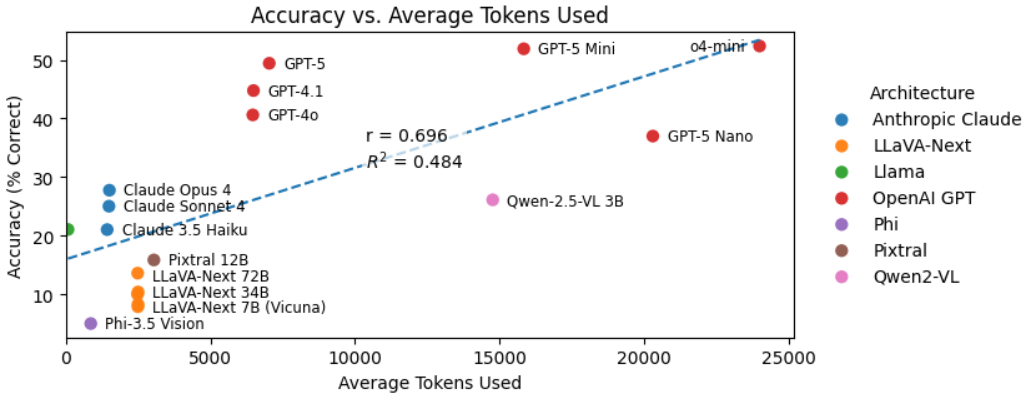


Figure 7: MapQA Benchmark performance compared to tokens used. Input tokens, chain-of-thought tokens, and output tokens are included in the total. Averages are computed per model with each dot representing one model's average accuracy and average number of tokens used. $r = 0.6976$, $p - value = 0.00129$, $R^2 = 0.4867$

can integrate symbolic abstraction, spatial reasoning, and visual semantics more effectively. Closing the human–machine gap on map understanding is essential for advancing AI systems capable of supporting high-stakes applications in navigation, environmental science, urban planning, and beyond. MapQA thus serves as a foundation for the next generation of multimodal reasoning research, where genuine map literacy becomes a prerequisite for real-world competence.

7 Acknowledgments

This work was supported by the DARPA Mathematics for the DIscovery of ALgorithms and Architectures (DIAL) program, the DARPA Knowledge Management at Scale and Speed (KMASS) program, the NSF award 2124052, the MIT-IBM Watson AI Lab, the Department of the Air Force Artificial Intelligence Accelerator under Cooperative Agreement Number FA8750-19-2-1000, and the Air Force Office of Scientific Research (AFOSR) under award number FA9550-21-1-0399, and the Developing the Airmen We Need–Education (DAWN-ED) PhD fellowship.

8 Disclaimer

The views expressed in this article are those of the author and do not reflect the official policy or position of the Department of the Space Force, Department of Defense, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. At the time of publication, Christian Arnold is an active duty officer of the United States Space Force.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024.

Meta AI. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models, September 2024. URL https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/. Meta AI Blog.

Anthropic. Introducing computer use, a new claude 3.5 sonnet, and claude 3.5 haiku. https://www.anthropic.com/news/3-5-models-and-computer-use, October 2024.

Anthropic. Introducing claude 4. https://www.anthropic.com/news/claude-4, May 2025.

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

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

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. Tomayto, tomahto: Beyond token-level answer equivalence for question answering evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 291–305, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.20. URL https://aclanthology.org/2022.emnlp-main.20/.

Roberto Casati. *The Cognitive Life of Maps*. The MIT Press, Cambridge, MA, 2023. doi: 10.7551/mitpress/14864.001.0001. URL https://direct.mit.edu/books/oa-monograph/5785/The-Cognitive-Life-of-Maps.

Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8748–8757, 2019.

Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. Mapqa: A dataset for question answering on choropleth maps. *arXiv preprint arXiv:2211.08545*, 2022.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *arXiv preprint arXiv:2105.14517*, 2021a.

Jifan Chen, Eunsol Choi, and Greg Durrett. Can NLI models verify QA systems' predictions? In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3841–3854, Punta Cana, Dominican Republic, November 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.324. URL https://aclanthology.org/2021.findings-emnlp.324/.

David Rumsey Map Collection. David rumsey historical map collection, 2025. URL https://www.davidrumsey.com.

Steve Fick and Robert Hijmans. Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37, 05 2017. doi: 10.1002/joc.5086.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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*, pp. 6904–6913, 2017.

M. C. Hansen, P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend. High-resolution global maps of 21st-century forest cover change. *Science*, 342(6160):850–853, 2013. doi: 10.1126/science.1244693. URL https://www.science.org/doi/abs/10.1126/science.1244693.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7856–7870, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.619. URL https://aclanthology.org/2021.emnlp-main.619/.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023) — Long Papers*, pp. 5591–5606, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.307. URL https://aclanthology.org/2023.acl-long.307/.

Sergios-Anestis Kefalidis, Dharmen Punjani, Eleni Tsalapati, Konstantinos Plas, Mariangela Pollali, Michail Mitsios, Myrto Tsokanaridou, Manolis Koubarakis, and Pierre Maret. Benchmarking geospatial question answering engines using the dataset geoquestions1089. In *The Semantic Web – ISWC 2023: 22nd International Semantic Web Conference, Athens, Greece, November 6–10, 2023, Proceedings, Part II*, pp. 266–284, Berlin, Heidelberg, 2023. Springer-Verlag. ISBN 978-3-031-47242-8. doi: 10.1007/978-3-031-47243-5_15. URL https://doi.org/10.1007/978-3-031-47243-5_15.

Nathaniel Vaughn Kelso and Tom Patterson. Natural earth vector. *Cartographic Perspectives*, 64:45–50, Sep. 2009. doi: 10.14714/CP64.148. URL https://cartographicperspectives.org/index.php/journal/article/view/cp64-kelso-patterson.

Tanja Kramm, Nicodemus Nyamari, Vincent Moseti, Annika Klee, Leon Vehlken, David M. Anderson, Christina Bogner, and Georg Bareth. Deep learning-based extraction of kenya's historical road network from topographic maps. *Scientific Data*, 12:1149, 2025. doi: 10.1038/s41597-025-05442-6. URL https://www.nature.com/articles/s41597-025-05442-6.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Joongbo Shin, and Kyomin Jung. KPQA: A metric for generative question answering using keyphrase weights. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2105–2115, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.170. URL https://aclanthology.org/2021.naacl-main.170/.

Zekun Li, Malcolm Grossman, Mihir Kulkarni, Muhao Chen, Yao-Yi Chiang, et al. Mapqa: Open-domain geospatial question answering on map data. *arXiv preprint arXiv:2503.07871*, 2025.

Library of Congress. Library of congress digital maps collection, 2025. URL https://www.loc.gov/maps/.

Elri Liebenberg, Imre Josef Demhardt, and Soetkin Vervust (eds.). *History of Military Cartography*. Lecture Notes in Geoinformation and Cartography. Springer International Publishing, 2016. doi: 10.1007/978-3-319-25244-5. URL https://link.springer.com/book/10.1007/978-3-319-25244-5.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566, 2020.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.

Philip Marthew. The power of cartography: How maps shape our understanding of the world. *Geoinfor Geostat: An Overview*, 11(1), 2023. doi: 10.4172/2327-4581.100032. URL https://www.scitechnol.com/peer-review/the-power-of-cartography-how-maps-shape-our-understanding-of-the-world-CVBX.php?article_id=21546.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

David Mayo, Jesse Cummings, Xinyu Lin, Dan Gutfreund, Boris Katz, and Andrei Barbu. How hard are computer vision datasets? calibrating dataset difficulty to viewing time. *Advances in Neural Information Processing Systems*, 36:11008–11036, 2023.

Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena. *arXiv preprint arXiv:2406.07545*, 2024.

Yelena Nakhimovsky, Andrew T. Miller, Tom Dimopoulos, and Michael Siliski. Behind the scenes of google maps navigation: Enabling actionable user feedback at scale. In *CHI 2010 Extended Abstracts on Human Factors in Computing Systems*, pp. 3763–3768, New York, NY, 2010. URL http://doi.acm.org/10.1145/1753846.1754052.

OpenAI. o4-mini model — openai api. URL https://developers.openai.com/api/docs/models/o4-mini.

OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, May 2024.

OpenAI. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/, April 2025a.

OpenAI. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/, August 2025b.

OpenAI. Gpt-5-nano. https://platform.openai.com/docs/models/gpt-5-nano, 2025c.

OpenStreetMap contributors. Openstreetmap, 2017. URL https://www.openstreetmap.org.

Dharmen Punjani, Kuldeep Singh, Andreas Both, Manolis Koubarakis, Iosif Angelidis, Konstantina Bereta, Themis Beris, Dimitris Bilidas, Theofilos Ioannidis, Nikolaos Karalis, et al. Template-based question answering over linked geospatial data. In *Proceedings of the 12th workshop on geographic information retrieval*, pp. 1–10, 2018.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018.

Chenglei Si, Chen Zhao, and Jordan Boyd-Graber. What's in a name? answer equivalence for open-domain question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic, 2021.

Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. Evaluation metrics in the era of GPT-4: Reliably evaluating large language models on sequence to sequence tasks. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8776–8788, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.543. URL https://aclanthology.org/2023.emnlp-main.543/.

Oleg Sychev. Open-answer question with regular expression templates and string completion hinting. *Software Impacts*, 17:100539, 2023. ISSN 2665-9638. doi: https://doi.org/10.1016/j.simpa.2023.100539. URL https://www.sciencedirect.com/science/article/pii/S2665963823000763.

José Teixeira and A. Teodoro. Remote sensing and gis applications in earth and environmental systems sciences. *SN Applied Sciences*, 3(1), 2021. doi: 10.1007/s42452-021-04855-3. URL https://link.springer.com/article/10.1007/s42452-021-04855-3.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578, June 2024.

U.S. Geological Survey. Landsat collections data archive. https://www.usgs.gov/landsat-missions/landsat-collections, 2025a.

U.S. Geological Survey. The national map: Geospatial data and services. https://www.usgs.gov/core-science-systems/national-geospatial-program/national-map, 2025b.

Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36:26650–26685, 2023.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*, 2023.

A DETAILED RESULTS

model	Full Set	Hand Annotated		HITL Generated		Multi Image		Single Image	
	Acc	Acc	Change	Acc	Change	Acc	Change	Acc	Change
GPT-o4-mini	0.52	0.53	(0.7%)	0.60	(7.5%)	0.28	(-24.3%)	0.58	(5.8%)
GPT-5 Mini	0.52	0.50	(-2.0%)	0.59	(6.8%)	0.32	(-19.5%)	0.57	(4.7%)
GPT-5	0.49	0.52	(2.3%)	0.55	(5.5%)	0.30	(-19.7%)	0.54	(4.7%)
GPT-4.1	0.45	0.47	(2.1%)	0.52	(7.0%)	0.21	(-24.2%)	0.51	(5.8%)
GPT-4o	0.41	0.41	(0.2%)	0.48	(7.5%)	0.17	(-24.0%)	0.46	(5.7%)
GPT-5 Nano	0.37	0.35	(-2.1%)	0.44	(7.1%)	0.17	(-20.1%)	0.42	(4.8%)
Claude Opus 4	0.28	0.36	(7.9%)	0.31	(3.0%)	0.10	(-17.6%)	0.32	(4.2%)
Qwen2.5-VL 3B	0.26	0.23	(-3.4%)	0.33	(6.5%)	0.09	(-17.1%)	0.30	(4.1%)
Claude Sonnet 4	0.25	0.34	(9.2%)	0.27	(2.4%)	0.08	(-17.0%)	0.29	(4.1%)
Claude 3.5 Haiku	0.22	0.32	(9.6%)	0.24	(1.3%)	0.09	(-13.6%)	0.26	(3.5%)
Llama-3 11B	0.21	0.25	(4.0%)	0.24	(2.7%)	0.09	(-12.5%)	0.24	(3.0%)
Pixtral 12B	0.16	0.24	(7.8%)	0.16	(0.5%)	0.06	(-9.8%)	0.18	(2.3%)
LLaVA-Next 72B	0.14	0.20	(6.3%)	0.16	(2.0%)	0.01	(-12.7%)	0.17	(3.0%)
LLaVA-Next 34B	0.10	0.15	(4.9%)	0.12	(1.3%)	0.01	(-9.1%)	0.13	(2.2%)
LN 7B (Mistral)	0.10	0.13	(3.0%)	0.11	(1.3%)	0.03	(-7.1%)	0.12	(1.7%)
LN 7B (Vicuna)	0.08	0.14	(5.4%)	0.09	(0.5%)	0.01	(-7.2%)	0.10	(1.7%)
LN 13B (Vicuna)	0.08	0.10	(2.6%)	0.09	(1.5%)	0.01	(-7.2%)	0.10	(1.7%)
Phi-3.5 Vision	0.05	0.04	(-0.9%)	0.06	(0.7%)	0.04	(-1.3%)	0.05	(0.3%)

Table 3: MapQA Subcategory Accuracy (Part 1). LlaVa-Next is abbreviated as LN for spacing. Human-in-the-Loop is abbreviated as HITL.

model	Military		Natural World		Urban		Aviation	
	Acc	Change	Acc	Change	Acc	Change	Acc	Change
GPT-o4-mini	0.64	(11.2%)	0.51	(-1.4%)	0.64	(11.1%)	0.30	(-22.4%)
GPT-5 Mini	0.61	(9.6%)	0.52	(0.2%)	0.57	(5.1%)	0.33	(-19.1%)
GPT-5	0.59	(10.0%)	0.45	(-4.5%)	0.61	(11.6%)	0.32	(-17.5%)
GPT-4.1	0.57	(12.0%)	0.42	(-2.7%)	0.52	(6.9%)	0.23	(-21.7%)
GPT-4o	0.52	(11.7%)	0.38	(-2.2%)	0.46	(5.7%)	0.19	(-21.2%)
GPT-5 Nano	0.45	(7.7%)	0.41	(3.9%)	0.42	(4.8%)	0.18	(-18.9%)
Claude Opus 4	0.34	(6.0%)	0.27	(-1.1%)	0.47	(19.4%)	0.13	(-14.8%)
Qwen2.5-VL 3B	0.34	(8.3%)	0.27	(0.4%)	0.33	(6.7%)	0.09	(-17.4%)
Claude Sonnet 4	0.31	(6.2%)	0.23	(-1.6%)	0.44	(18.8%)	0.10	(-14.6%)
Claude 3.5 Haiku	0.28	(5.4%)	0.21	(-0.9%)	0.39	(17.1%)	0.10	(-12.1%)
Llama-3 11B	0.29	(7.6%)	0.15	(-5.7%)	0.33	(11.7%)	0.09	(-12.0%)
Pixtral 12B	0.22	(6.5%)	0.08	(-7.6%)	0.33	(16.9%)	0.06	(-9.4%)
LLaVA-Next 72B	0.19	(5.4%)	0.11	(-2.8%)	0.32	(18.0%)	0.02	(-11.7%)
LLaVA-Next 34B	0.14	(3.7%)	0.08	(-2.2%)	0.24	(13.8%)	0.02	(-8.1%)
LN 7B (Mistral)	0.13	(2.6%)	0.09	(-0.5%)	0.20	(10.5%)	0.03	(-6.9%)
LN 7B (Vicuna)	0.11	(2.8%)	0.06	(-2.7%)	0.24	(15.5%)	0.02	(-6.5%)
LN 13B (Vicuna)	0.11	(2.7%)	0.07	(-0.7%)	0.18	(10.2%)	0.01	(-6.7%)
Phi-3.5 Vision	0.05	(-0.3%)	0.08	(3.0%)	0.02	(-3.3%)	0.04	(-1.2%)

Table 4: MapQA Subcategory Accuracy (Part 2). LlaVa-Next is abbreviated as LN for spacing. Human-in-the-Loop is abbreviated as HITL.

B DETAILED BREAKDOWN OF MAP TYPES AND SUBCATEGORIES

	Map Types	Number
1	Natural World	153
2	Urban	155
3	Aviation	138
4	Military	132
Total:	**4 Map Types**	**578**

Table 5: Breakdown of Map types with corresponding image counts

	Subcategory	Number
1	Park Map	3
2	Postal Roads	2
3	Sectional Chart	29
4	Terminal Area Chart	28
5	Helicopter	18
6	Strategic	82
7	Topographic	63
8	Transportation	61
9	Linguistic Map	7
10	Tactical	50
11	Soil Map	34
12	Aviation Maps	14
13	City Map	39
14	Hydrologic Map	2
15	Nautical Chart	49
16	Zoning Map	11
17	Cadastral	8
18	Insurance Map	18
19	Distribution Map	4
20	Survey Map	2
21	Aviation Charts	5
22	Administrative	3
23	Fantasy Map	2
24	Historic	44
Total:	**24 subcategories**	**578**

Table 6: Breakdown of subcategories with corresponding image counts

C GROUPED DIFFICULTY ACCURACY

In fig. 8, we compare the average model accuracy grouped by difficulty based on user response time. As demonstrated by Mayo et al. (2023), it is common for questions that are easy for humans to also be easy for vision-enabled models. As expected, the Very Easy difficulty group shows the highest performance across all models, with GPTo4-mini and GPT5-mini exceeding 70% accuracy. As difficulty (inferred by human response time) increased, the performance of each model decreased with the exception of the Medium difficulty group. More analysis is needed to determine the possible cause of this unexpected outcome.

D LACK OF GEMINI TESTING

We attempted to run inference using Gemini, but after processing 1k questions we encountered a file storage quota issue with our Google Cloud account, despite not using any file storage. We have an open ticket with support and hope to complete the evaluation of Gemini once it is resolved.

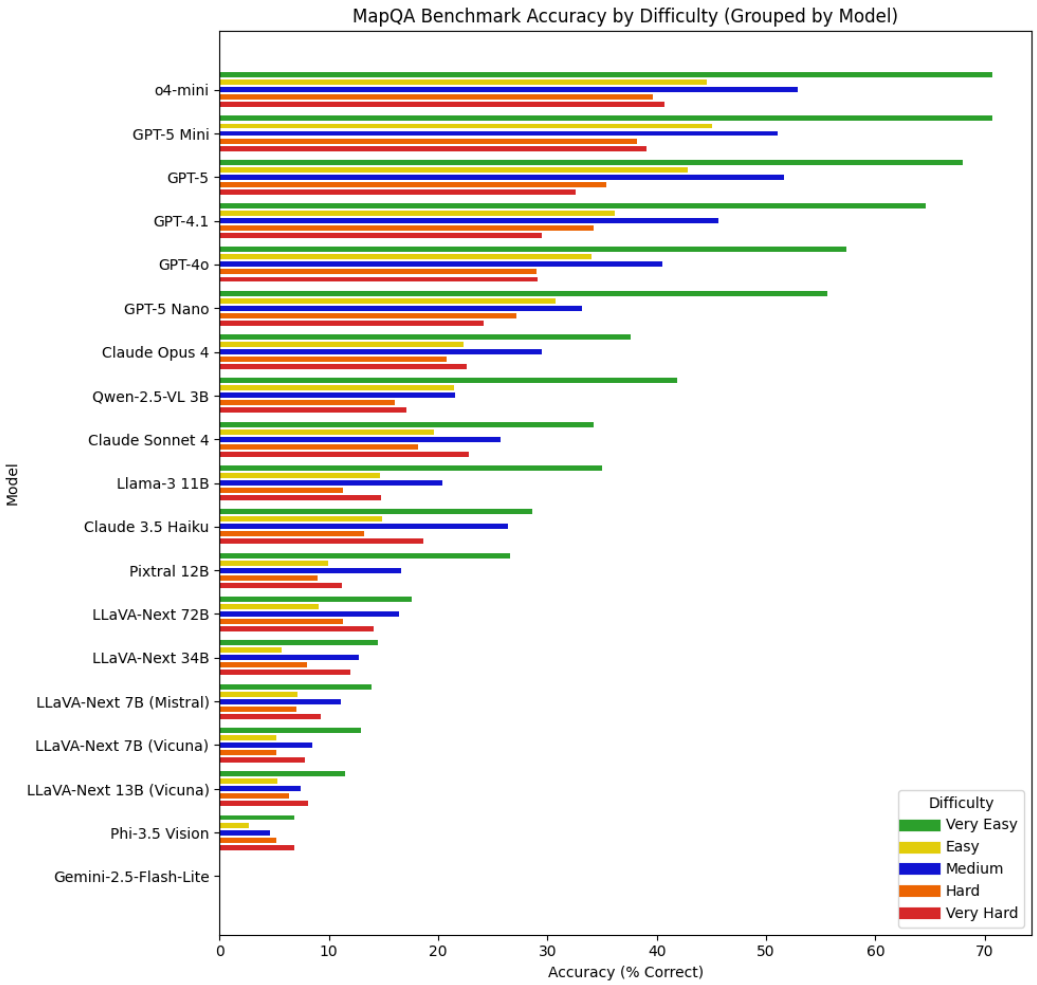


Figure 8: MapQA Benchmark performance for each difficulty.

E USE OF LLMS

An LLM and other writing tools were used to detect grammar mistakes and typos during the last stages of writing. An LLM was also used to aid in the writing of python scripts and functional prototyping.

F ETHICS

The MapQA benchmark was developed without collecting or storing any personally identifiable information. All map images used are sourced from government archives in the public domain. Human annotators who participated in the curation process were recruited as part of the research team or trained collaborators, and their contributions did not involve sensitive personal data.

Because MapQA focuses on evaluating visual reasoning over maps, potential risks stem from downstream applications of models trained or tested on this benchmark. Misinterpretation of maps by automated systems could have harmful consequences in navigation, disaster response, environmental monitoring, or urban planning. At present, MapQA is released as an evaluation-first benchmark, ensuring that models are rigorously tested before being deployed in such contexts. In the future, we anticipate that MapQA may also serve as a training resource, provided this is done responsibly with attention to ethical safeguards and careful consideration of downstream risks.

G REPRODUCIBILITY

We have taken multiple steps to ensure reproducibility. The MapQA benchmark itself is released in full as part of the supplementary materials (single file), including all image file links, questions, and reference answers. The images are not directly included as they are over 5GB in total, exceeding 100MB limit for OpenReview supplementary material submissions. Section 3 (The MapQA Benchmark) provides a detailed description of dataset statistics, splits, and the curation pipeline. Evaluation was performed with FuseKit, a unified model testbench developed by the coauthors for rapid evaluation of open source and proprietary API models: https://github.com/Scuwr/FuseKit.

Our evaluation protocol, including the LLM-as-a-Judge setup, calibration procedure, and system prompts, is described in Section 4 (Experiments) and Figure 4. Model inference configurations (devices, sharding strategies, and prompts) are documented in the main text. Baseline accuracies for all tested models are reported in Figures 5, 6, 7, and 6.

To replicate results, researchers need only (i) the released benchmark file, (ii) an evaluation script and images that will be provided in final submission, and (iii) access to the models under test. We note that hardware requirements differ by model size; these details are specified in Section 4.1 (Baselines). Together, these materials enable independent verification of dataset statistics, evaluation results, and analysis.

H ANNOTATION INSTRUCTIONS

We will be evaluating how different AIs interact with maps. You will be given a map and a question to answer about it. The questions concern topics such as physical features in the map, information from pre-existing map knowledge like latitude lines, and information from charts.

Your responsibility is to answer the question using the map. Some but not all of these questions were generated using AI models. Because of this, some questions might be bad; they might be vague or ambiguous, unanswerable, or too difficult. We have provided examples in the Appendix.

You will be randomly assigned datasets based on time of completion, and the numbering might not be in order.

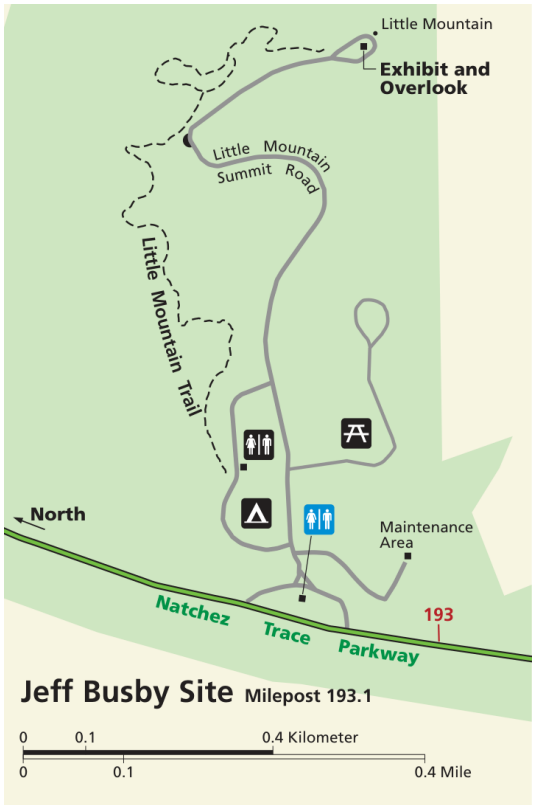
Your Responsibility

1. **Answer the question as best as you can**. Answers should be provided regardless of whether it is a good question or not. **Please feel free to use Google and/or Google/Apple Maps to locate landmarks, understand terms, or to help orient yourself on the map.** Please note that if the image appears unclear, going into full screen will open a higher-resolution version.
   A good answer SHOULD:
   - Be factual and clear.
   - Rely on the map, not pre-existing knowledge of geography (although it may rely on existing knowledge of map conventions).
2. **Determine if it's a good or bad question**
   A good question SHOULD:
   - Be answerable with just the map, and an ordinary person's knowledge of how to interpret maps.
   - Have a single clear answer — if there are multiple correct answers, it is not a good question.
   - Sometimes questions will not have an answer; simply write "None" or "There is no 'X' on the map."

   A good question should NOT:
   - Require you to count values greater than 10.
   - Be very difficult. It may take a while to find the map items referred to in the question, but once you do, answering it should be simple.

If it's a bad question, please STILL make your best attempt to answer the question. We have provided a box you can check which will reveal a field where you can input your explanation for why it is a poor question. Please fill out this part **ONLY** if you think it is a bad question.

Examples of Bad Questions



Bad Question (Too Vague/Ambiguous): What color is the bathroom sign?

Explanation: There are two signs signifying bathrooms and both have different colors: blue and black.

Bad Question (Unanswerable with Map): Where is Chicago relative to the Exhibit?

Explanation: Chicago is neither directly on the map nor referenced on the map and thus impossible to answer without additional information.

Bad Question (Too Difficult): How many dashes are used to label the Little Mountain Trail?

Explanation: Requires highly precise counting. A good rule of thumb is that anything with a count higher than 8 is "unfair".