GeoChain: Multimodal Chain-of-Thought for Geographic Reasoning

Sahiti Yerramilli[*]
Google
sahitiy@google.com

Nilay Pande[*]
Waymo
nilayp@waymo.com

Rynaa Grover[*]
Google
rynaa@google.com

Jayant Sravan Tamarapalli[*]
Google
jayantsravan@google.com

Abstract

This paper introduces GeoChain, a large-scale benchmark for evaluating step-by-step geographic reasoning in multimodal large language models (MLLMs). Leveraging 1.46 million Mapillary street-level images, GeoChain pairs each image with a 21-step chain-of-thought (CoT) question sequence (over 30 million Q&A pairs). These sequences guide models from coarse attributes to fine-grained localization across four reasoning categories - visual, spatial, cultural, and precise geolocation - annotated by difficulty. Images are also enriched with semantic segmentation (150 classes) and a visual locatability score. Our benchmarking of frontier MLLMs on a diverse 2,088-image subset reveals consistent challenges: models frequently exhibit weaknesses in visual grounding, display erratic reasoning, and struggle to achieve accurate localization, especially as the reasoning complexity escalates. GeoChain offers a robust diagnostic methodology, critical for fostering significant advancements in complex geographic reasoning within MLLMs.

Code: GeoChain on GitHub

Dataset: GeoChain on Hugging Face

1 Introduction

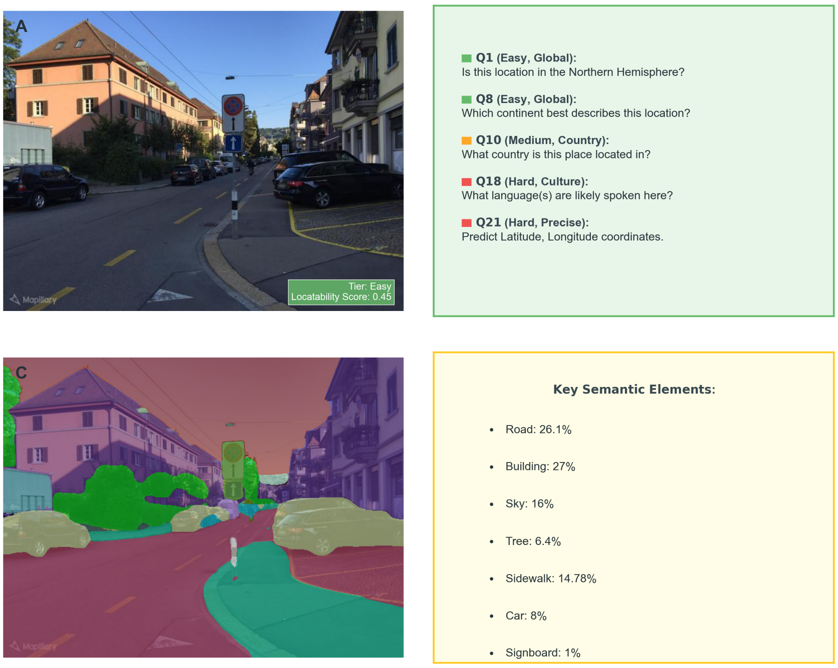
As large vision-language models (VLMs) continue to make rapid progress on general visual question answering and captioning tasks (Team et al., 2024; OpenAI et al., 2024; Wang et al., 2024; Dai et al., 2023), their capacity for structured geographic reasoning remains underexplored. The ability to infer a location from visual cues -such as terrain, signage, vehicles, or architecture - considered alongside spatial and cultural knowledge, is crucial for real-world applications like remote sensing, disaster response, and autonomous navigation. More broadly, geographic localization serves as a testbed for grounded intelligence, requiring models to reason over subtle visual features, incorporate world knowledge, and disambiguate locations that may be visually similar. Despite this, existing benchmarks rarely probe the kind of step-by-step reasoning that such tasks demand.

We introduce GeoChain, a novel multimodal benchmark for evaluating structured geographic reasoning in large language models (MLLMs). As depicted in Figure 1, each GeoChain sample features a street-level image from the Mapillary dataset (Warburg et al., 2020) paired with a 21-step chain-of-thought (CoT) question sequence. These sequences progressively guide models from coarse inferences, such as hemisphere or continent, to fine-grained predictions like city, latitude, and longitude. The complete GeoChain framework comprises 1.46 million images, each with this 21-question CoT structure, yielding over 30 million question-answer pairs. Questions span four core reasoning categories: visual cues, spatial localization, cultural inference, and precise geolocation, all annotated with difficulty levels for granular evaluation. This curriculum-style structure offers vital diagnostic insights into where and why models fail across reasoning stages, moving beyond sole reliance on final predictions.

To facilitate focused evaluations, we curated GeoChain Test-Mini, a diverse and challenging subset. This curation process leverages a locatability score, adapted from GeoReasoner (Li et al., 2024) and computed using features from a pretrained MaskFormer model (Cheng et al., 2021). This score quantifies the visual identifiability of a location from a single image, allowing us to stratify GeoChain Test-Mini into Easy, Medium, and Hard tiers based on thresholds in the 0.12-0.6 range. The resulting GeoChain Test-Mini contains 2,088 carefully selected images, designed to offer a representative yet manageable scale for robust MLLM assessment.

[*] Equal Contribution.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 23624–23639
November 4-9, 2025



🟩 **Q1 (Easy, Global):**
Is this location in the Northern Hemisphere?

🟩 **Q8 (Easy, Global):**
Which continent best describes this location?

🟧 **Q10 (Medium, Country):**
What country is this place located in?

🟥 **Q18 (Hard, Culture):**
What language(s) are likely spoken here?

🟥 **Q21 (Hard, Precise):**
Predict Latitude, Longitude coordinates.

Key Semantic Elements:

- Road: 26.1%
- Building: 27%
- Sky: 16%
- Tree: 6.4%
- Sidewalk: 14.78%
- Car: 8%
- Signboard: 1%

Figure 1: Components of a GeoChain instance: **(Top-Left)** Easy Mapillary Street-Level Sequences (MSLS) image with locatability score of 0.45. **(Top-Right)** Example chain-of-thought questions with difficulty indicators. **(Bottom-Left)** Derived semantic segmentation map. **(Bottom-Right)** Extracted key semantic labels. Together, these elements enable step-by-step diagnostic evaluation of geographic reasoning.

Our main contributions are as follows:

- **GeoChain Benchmark Framework:** A novel benchmark for evaluating step-by-step MLLM geographic reasoning, derived from 1.46 million Mapillary street-level images and generating over 30 million Q&A pairs through 21-step chain-of-thought questions, all structured across diverse reasoning categories and difficulty levels.

- **Rich Augmentation & Curated Evaluation Set:** A methodology for enhancing images with semantic labels (150 classes) and a human-inspired locatability score for difficulty stratification, culminating in the GeoChain Test-Mini: a quality-controlled 2088-image evaluation set; the resulting rich semantic metadata also offers a valuable resource for broader community research and future investigations.

- **Comprehensive MLLM Benchmarking & Analysis:** Evaluation of leading MLLMs on GeoChain Test-Mini, providing detailed insights into their geographic reasoning capabilities, performance variations, and common failure modes.

2 Related Work

2.1 Image-Based Geolocation

Early work in visual geolocation predominantly focused on matching query images to large, geo-tagged image databases, often aiming for direct coordinate prediction. For instance, Im2GPS (Hays and Efros, 2008) pioneered retrieving locations by comparing against a massive photograph dataset. Later, deep learning significantly advanced the field; PlaNet (Weyand et al., 2016) utilized convolutional neural networks (CNNs) for global location prediction, and architectures like NetVLAD (Arandjelovic et al., 2016) learned robust image representations for effective place recognition, improving upon earlier retrieval methods. Other approaches, such as those focusing on urban or cross-view settings (Tian et al., 2017), further specialized these techniques. GeoChain diverges from these paradigms, which primarily target endpoint localization accuracy or image retrieval. Instead, it introduces a structured multimodal reasoning benchmark where models must articulate a 21-step chain-of-thought (CoT) sequence of answers to geographically relevant questions, thereby enabling finer-grained diagnostic insight into their internal reasoning processes.

2.2 Multimodal Geographic Reasoning and Benchmarks

More recent efforts have begun to integrate visual understanding with language-based reasoning for complex geographic tasks. GeoReasoner (Li et al., 2024), for example, introduced a fine-tuning strategy for MLLMs using human gameplay traces, primarily to improve final location prediction by modeling human-like inference. Similarly, other recent studies (Pramanik et al., 2024; Yang et al., 2024) also concentrate on predicting precise latitude and longitude. GeoComp (Song et al., 2025) presents a large-scale dataset of geolocation gameplay data, emphasizing step-wise reasoning rooted in real human gameplay that often involves external metadata, active exploration, and dynamic information gathering. While these approaches offer valuable insights into human-like inference and gameplay dynamics, GeoChain's contribution is complementary. It does not involve model fine-tuning or rely on gameplay trajectories. Instead, GeoChain employs a fully static, image-grounded evaluation framework: each sample consists of a single image paired with its fixed CoT question sequence, standardized across the entire benchmark. This design facilitates direct and controlled benchmarking of different models' inherent reasoning capabilities under uniform conditions, distinct from evaluating exploratory strategies or the ability to process dynamic data.

Other benchmarks, such as GAEA (Campos et al., 2025), generate diverse conversational questions from detailed, place-specific metadata like OpenStreetMap attributes. While this can create rich contextual queries, it introduces challenges related to the temporal stability of dynamic data (e.g., changes in urban landscape) and complicates fair, apples-to-apples model comparisons due to non-uniform question sets. Consequently, disentangling model reasoning failures from idiosyncratic question characteristics becomes difficult. GeoChain mitigates these issues by grounding its standardized questions in more enduring visual semantics such as the presence of characteristic vegetation, architectural styles, or road infrastructure, often identifiable through image segmentation and stable general geographic facts. This focus ensures the evaluation centers on the consistency of the reasoning process itself.

Furthermore, existing geospatial benchmarks like GEO-Bench (Lacoste et al., 2023) primarily target remote sensing applications, offering valuable tools for Earth monitoring with satellite imagery and tasks such as classification or segmentation. In contrast, GeoChain specifically addresses agent-level geographic reasoning from high-resolution, ground-level imagery, emphasizing natural-language understanding of spatial, cultural, and visual cues directly perceivable in such environments.

2.3 Mapillary Street-Level Sequences Dataset

GeoChain is built upon the Mapillary Street-Level Sequences (MSLS) dataset (Warburg et al., 2020), a large-scale, crowd-sourced collection of diverse, geo-tagged street-level images. MSLS's global coverage, with data from numerous cities worldwide reflecting the breadth of the MSLS ecosystem, and its varied capture conditions (diverse cameras, viewpoints, seasons, times of day) make it an ideal foundation for a benchmark aiming to evaluate generalizable geographic reasoning.

3 GeoChain Benchmark Construction

The GeoChain benchmark is constructed by augmenting the Mapillary Street-Level Sequences (MSLS) dataset (Warburg et al., 2020). MSLS provides a diverse collection of geo-tagged street-level imagery (approximately 1.4 million images in its full extent, with a geographical distribution across numerous cities as illustrated in Figure 2), crucial for developing and evaluating geographic localization models. However, to facilitate fine-grained, step-by-step reasoning, we introduce several layers of annotation and metadata. Our contributions enhance the MSLS dataset in three primary ways: semantic class labeling, locatability score computation, and the design of a structured chain-of-thought question battery. These augmentations, followed by a careful test set curation process, collectively enable a more nuanced evaluation of multimodal models' geographic reasoning capabilities.

3.1 Semantic Class Labeling

To ground visual reasoning in explicit semantic content, each image in our benchmark is augmented with a semantic segmentation map. This map provides a detailed understanding of the scene's composition by identifying various objects and environmental features. We employed MaskFormer (Cheng et al., 2021), a state-of-the-art transformer-based architecture for semantic segmentation. Specifically, we utilized a MaskFormer

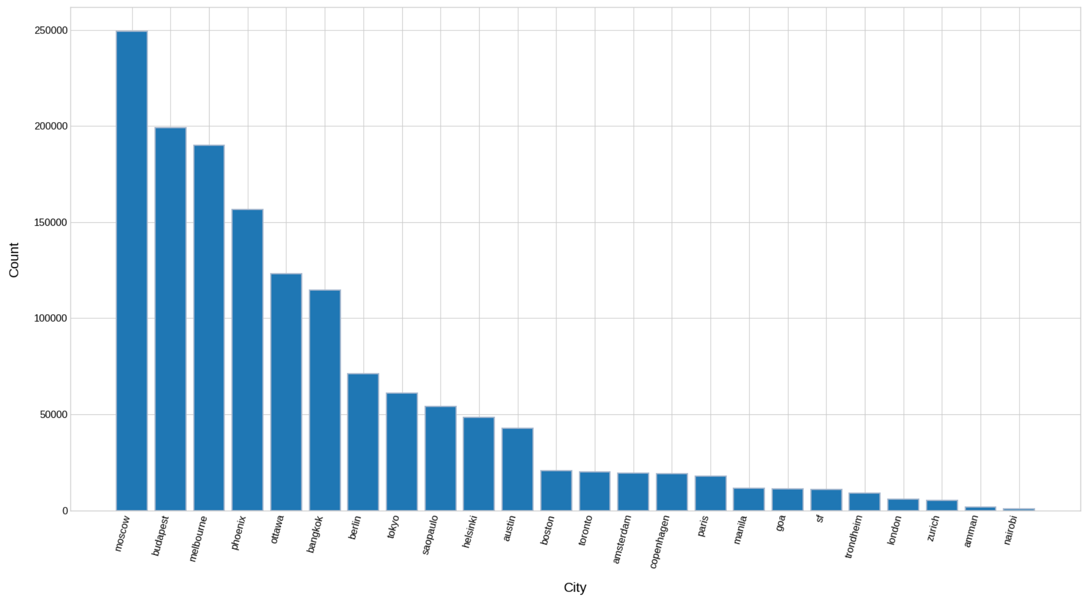


Figure 2: Count of images per city, illustrating the city distribution within the GeoChain dataset.

model pre-trained on the ADE20K dataset (Zhou et al., 2017), which offers a rich label set of 150 distinct classes, encompassing a wide array of objects, environmental elements (e.g., “tree”, “sky”, “road”), and architectural features (e.g., “building”, “window”, “door”).

From the segmentation map, we calculate how much of the image is covered by each category. We do this by working out the percentage of the image’s total area that each specific category takes up. For example, we might find that ‘sky’ covers 30% of an image, and ‘road’ covers 15%. This measurement of what’s in the scene, and how much of it there is, then helps us create the correct answers for many of the visual questions in our benchmark.

3.2 Locatability Score Computation

To systematically assess model performance across varying levels of visual ambiguity, we compute a *locatability score* for every image considered for the GeoChain benchmark. This score, ranging from 0 to 1, quantifies how visually identifiable a location is likely to be, with higher scores indicating more distinct and easily locatable scenes. Our methodology for calculating this score is adopted from (Li et al., 2024). The distribution of these computed locatability scores across the considered images is shown in Figure 3.

The core idea behind this score is to leverage common visual cues that humans, particularly proficient GeoGuessr players (GeoGuessr, 2013), rely on for geolocalization. The process involves several steps:

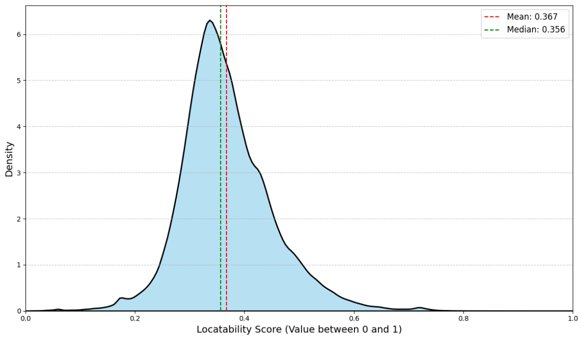


Figure 3: Distribution of Locatability Scores in GeoChain

1. **Identification of Cues:** A set of cues frequently used by GeoGuessr players (e.g., “houses in central Chile are more likely to have terracotta tiled roofs”) is established.

2. **Cue-to-Class Similarity:** The semantic similarity between these cues and the 150 class labels produced by the MaskFormer model (as described in Section 3.1) is computed. This typically involves using text embeddings to represent both the cues and the class labels, followed by a similarity measure (e.g., cosine similarity).

3. **Class Weight Derivation:** The similarities are aggregated across all cues for each class and then subjected to min-max normalization to derive a set of weights w_c for each class c. These weights reflect the importance of each visual class for

geolocalization.

4. **Weighted Score Aggregation:** The final locatability score for an image is computed as a weighted sum of the percentage areas of the classes present in the image.

This locatability score is then used to stratify the images within the GeoChain test set into three distinct tiers: **Hard** if score $\in [0.12, 0.22)$, **Medium** if score $\in [0.22, 0.45)$, and **Easy** if score $\in [0.45, 0.6)$. This stratification, based purely on visual cues inherent in the imagery, allows for a more granular analysis of model performance and helps identify specific weaknesses in reasoning about visually challenging environments.

3.3 Chain-of-Thought Question Design

A central component of GeoChain is a carefully designed sequence of 21 questions that guide the model through a step-by-step reasoning process, from coarse-grained observations to fine-grained localization. This chain-of-thought (CoT) approach aims to mimic a structured human-like deduction process. The questions are ordered such that earlier questions elicit information or focus attention on attributes that can be instrumental in answering subsequent, more complex questions.

The full list of 21 questions, along with their rank, assigned difficulty (Easy, Medium, Hard), question type (Binary, Multiclass, Free-text), and question category (e.g., Culture/Infrastructure, Geo Localization, Terrain/Environment), is provided in Appendix A.1.1. The difficulty annotation (Easy, Medium, Hard) for each question reflects the anticipated challenge of answering that specific question in isolation, based on the type of information required.

The question set is designed to be static across all data points in the benchmark. This uniformity ensures a consistent evaluation framework, allowing for direct, apples-to-apples comparisons of different models' reasoning capabilities. The questions cover diverse aspects:

- **Visual Object/Attribute Presence:** Some questions directly query the presence of specific objects or attributes identifiable from the image (e.g., "Do you see any boats or ships?"). Ground truth for these questions is primarily derived from the semantic class labels extracted via the MaskFormer model (Section 3.1). For instance, if the class "boat" occupies a non-zero percentage of the image, the answer would be affirmative.
- **Inferential and Contextual Knowledge:** Other questions require more derivative reasoning or contextual knowledge beyond direct object identification (e.g., "Is this place near a coast?", "What side of the road do vehicles drive on here?"). The MSLS dataset encompasses images from 24 distinct cities globally. For images originating from these locations, we manually curated ground truth answers for city-level attributes or environmental characteristics (e.g., predominant architectural styles, typical climate indicators) that apply broadly to the image's geographic area.
- **Progressive Localization:** The sequence progresses from general observations (e.g., hemisphere, continent) to specific details (e.g., country, city, precise latitude and longitude coordinates).

The question types include binary (Yes/No), multiclass (selection from a predefined set of options), and free-text (open-ended answers, such as country name or coordinates). This variety tests different aspects of a model's understanding and generation capabilities.

The semantic segmentation labels generated in Section 3.1 were instrumental in constructing several questions that directly probe the visual understanding capabilities of the models. Beyond their use in the current benchmark, this rich semantic metadata, now part of GeoChain, offers a valuable resource for the community. It can be leveraged to design new questions aimed at further investigating specific aspects of model behavior, such as tendencies towards visual hallucination (Li et al., 2023) (Rohrbach et al., 2018) or the fine-grained ability to identify a wider array of objects. The insights derived from such extended evaluations can subsequently guide targeted improvements in model development.

By analyzing model performance across this structured chain of questions, GeoChain aims to provide deeper insights into the strengths and weaknesses of multimodal geographic reasoning systems.

3.4 Test Set Curation and Sampling Strategy

To create the "GeoChain Test-Mini" subset for focused evaluation, we prioritized stratification, visual quality, and diversity. We initially targeted

2100 images, stratified by locatability scores into 700 Easy, 700 Medium, and 700 Hard examples. A tiered, unique-sequence sampling strategy was employed: unique image sequences were randomly sampled first for the Hard tier, then for the Medium tier (from remaining unique sequences), and finally for the Easy tier, ensuring no sequence was reused across tiers. The underlying MSLS dataset exhibits a notable skew in its per-city image distribution (as highlighted by the overall dataset statistics in Figure 2). Consequently, to avoid introducing new biases that could arise from attempting to manually balance city representation or 'carefully' over/under-sample from specific locations, our approach was to randomly sample unique image sequences across all available cities within each defined locatability tier. These 2100 candidates underwent manual visual inspection, where 12 images with critical quality issues (e.g., excessive blur, poor exposure) were removed. This rigorous curation yielded a final Test-Mini set of 2088 high-quality, diverse, and appropriately challenging images.

4 Analysis

In this section, we evaluate the performance of frontier vision-language models: GPT-4.1, GPT-4.1-mini (OpenAI et al., 2024), Claude 3.7 Sonnet (Sonnet, 2025), Gemini 2.5 Flash (Google, 2025a), Gemini 2.5 Pro (Google, 2025b), Qwen 2.5 VL (Bai et al., 2025), Pixtral Large (Mistral AI, 2024), and Gemma 3 (Team et al., 2025) on the GeoChain "Test-Mini" benchmark, focusing on their ability to reason accurately and consistently across a structured 21-step geographic reasoning chain.

4.1 Evaluation Metrics

4.1.1 Haversine Distance

The final question in each GeoChain sequence (Question 21) requires the model to predict the geographic coordinates (latitude, longitude) of the depicted scene. To evaluate the accuracy of these specific predictions, we compute the *Haversine distance*: the shortest distance over the Earth's surface between the predicted and ground-truth coordinates, assuming a spherical Earth. A detailed explanation of the Haversine distance calculation is provided in Section A.2.

4.1.2 Pass Score

The **Pass Score** is computed as the average fraction of questions correctly answered across the full 21-step reasoning chain for each image. A prediction for any question is considered correct if it matches the ground-truth answer for that specific question, accounting for its type (e.g., exact match for free-text, class match for multiclass, or binary match). Crucially, for the final latitude and longitude prediction (Question 21), a response is deemed correct contributing to the Pass Score if its Haversine distance (as defined in Section 4.1.1) from the ground truth is less than 50km.

4.2 Overall Model Performance

Overall model performance (Table 1) offers nuanced insights into current MLLM geographic reasoning. The leading Gemini models exhibit specialized strengths: Gemini-2.5-pro excels in complex multi-step reasoning (pass score 81.84%), whereas Gemini-2.5-Flash achieves superior localization precision (445.24 km mean error), hinting at differing architectural or training optimizations. This divergence underscores that broad inferential ability and precise geolocalization are distinct skills, likely requiring separate optimization pathways rather than being monolithic capabilities. GPT-4.1 maintains a competitive position; however, the substantial localization inaccuracies of the other models underscore that robust geospatial grounding is a significant developmental hurdle, indicating a key area for advancement in MLLM capabilities.

The introduction of threshold-based localization accuracies - at City (< 25 km), Region (< 200 km), and Country (< 750 km) levels further refines this performance landscape. Gemini-2.5-pro's superior performance is reinforced by its top-tier City-level precision (59.38%). Complementing this, Gemini-2.5-Flash excels in broader accuracy, leading at both Region-level (70.02%) and Country-level (90.31%). GPT-4.1 also demonstrates notable strength in City-level performance (57.84%), surpassing Gemini-2.5-Flash in this specific high-precision context. The challenges in Claude 3.7 Sonnet, GPT 4.1 Mini, and other open source models are starkly emphasized by their profound difficulties at these finer scales. These granular metrics effectively highlight that achieving reliable, high-confidence city-level precision is a primary differentiator and a significant challenge across the evaluated MLLMs.

To provide a task-specific point of comparison, we also include GeoReasoner (Li et al., 2024), a contemporary VLM-based geolocation model. Since it only predicts country and city, we evaluate

Table 1: Overall model-level accuracy and localization metrics.

Model	Pass Score (%)	Mean Dist (km)	< 25 km (%)	< 200 km (%)	< 750 km (%)
Gemini-2.5-pro	**81.84**	489.51	**59.38**	69.95	88.51
Gemini-2.5-Flash	79.77	**445.24**	55.71	**70.02**	**90.31**
GPT-4.1	79.25	611.89	57.84	67.36	86.24
Claude 3.7 Sonnet	76.23	1289.04	40.34	47.07	73.31
GPT-4.1 Mini	70.42	1194.77	48.61	52.87	72.77
Qwen 2.5 VL (72B)	74.6	1067.0	41.7	49.0	75.0
Pixtral Large (124B)	68.1	1869.8	27.9	34.2	59.8
Gemma 3 (27B)	59.4	1235.5	34.4	43.9	72.0
GeoReasoner	—	1139.57*	4.76	—	39.49

*Distance is an approximation calculated from the predicted city's center.

it on the relevant metrics. Despite its task-specific design, its city-level accuracy is substantially lower than that of the top general-purpose models, and its mean distance error is high. This result reinforces the profound difficulty of the geographic reasoning task and validates the need for a granular, step-by-step benchmark like GeoChain to diagnose model failures.

4.3 Breakdown by Image Difficulty

Analyzing model performance by image difficulty (Table 2) reveals critical operational characteristics. As expected, 'Hard' images significantly challenge all models, leading to substantial increases in mean localization errors often exceeding 1000-2000 km for several models. The Gemini models consistently lead: Gemini-2.5-pro achieves top Pass Scores across all difficulties (e.g., 78.0% on Hard), while Gemini-2.5-Flash generally provides superior localization on 'Easy' and 'Medium' images (e.g., 188.45 km on Medium). Notably, Gemini-2.5-pro performs the best for localization precision on 'Hard' images (866.62 km), possibly where its stronger inferential capacity becomes decisive. An intriguing anomaly is the better localization by some models, like Gemini-2.5-Flash, on 'Medium' versus 'Easy' images, potentially due to bias towards certain cities in pre-training data. Furthermore, Claude 3.7 Sonnet's performance is particularly interesting: despite reasonable Pass Scores (e.g., 73.2% on Hard), its poor localization (2000.14 km on Hard) highlights a profound disconnect between understanding cues and grounding them spatially.

4.4 Breakdown by Question Category

Analyzing Pass Scores by question category (Table 3), informed by the benchmark's diverse question structures (e.g., visual queries versus free-text specific knowledge), reveals distinct performance strata. Foundational "Visual" questions, focusing on direct object presence (e.g., "Do you see any boats?"), yield universally high scores (all models >82%), suggesting robust basic visual grounding and low immediate hallucination, with Claude 3.7 Sonnet leading (92.8%). Similarly, "Terrain" identification is generally strong. In contrast, categories like "Geo Localization" and "Cultural" show mixed results; models likely handle simpler, coarse queries (e.g., continent identification) better than challenging free-text questions requiring specific knowledge (e.g., city/state names, language identification). Unsurprisingly, "Exact Loc" demanding precise latitude/longitude output—is definitively the most challenging category across all models. Within this landscape, Gemini-2.5-pro consistently excels, particularly in the more demanding categories like "Terrain" (87.4%), "Cultural" (77.9%), and "Exact Loc." (63.5%). GPT-4.1 also demonstrates strong performance, notably in "Geo Localization" (76.9%) and "Exact Loc." (61.5%). Claude 3.7 Sonnet's profile, with its excellent "Visual" scores but significantly weaker "Exact Loc." performance (51.0%), starkly illustrates a common theme: a disconnect between initial cue processing and final, precise geospatial grounding, which remains the primary MLLM hurdle.

4.4.1 Accuracy vs. Reasoning Depth

Figure 4 reveals a typical degradation pattern: All models perform well in the initial questions (1–9), which ask about visual or global cues such as ve-

Table 2: Performance by image difficulty. Accuracy (%) and Haversine distance (km) for each difficulty level.

Model	**Diff**	**Pass Score**	**M. Dist.**
Claude 3.7 Sonnet	Easy	77.2	885.86
	Medium	78.3	989.13
	Hard	73.2	2000.14
GPT-4.1 Mini	Easy	70.8	863.19
	Medium	73.2	827.78
	Hard	67.3	1910.44
GPT-4.1	Easy	79.3	357.36
	Medium	81.6	428.46
	Hard	76.8	1052.13
Gemini-2.5 Flash	Easy	80.5	**287.61**
	Medium	82.5	**188.45**
	Hard	76.3	873.78
Gemini-2.5 Pro	Easy	**83.3**	300.29
	Medium	**84.2**	304.32
	Hard	**78.0**	**866.62**
Qwen 2.5 VL	Easy	74.7	739.49
	Medium	77.1	853.91
	Hard	71.8	1636.54
Pixtral Large	Easy	67.9	1514.53
	Medium	70.1	1609.17
	Hard	66.1	2491.22
Gemma 3	Easy	60.3	889.06
	Medium	61.1	1056.79
	Hard	56.6	1766.11

hicles, hemisphere, or continent. These are relatively easy to infer on the basis of surface-level features. As the questions become more complex and semantically demanding, the accuracy drops sharply, especially at questions 10 and 17. These questions requiring nuanced interpretation of environmental and infrastructure signals. In particular, we observe a performance bump around questions 12–14. Despite appearing later in the sequence, these questions ask about relatively easy visual features (e.g., desert, hills, or city size). This reinforces the value of structuring questions not just by logical sequence but also by measured difficulty, allowing finer-grained diagnostics of model capability.

The final steps of the chain (questions 18–21) see the steepest drop in performance, as models are asked to predict language, administrative region, city name, and exact coordinates - tasks that require multi-modal reasoning, robust world knowledge, and low-level visual grounding. This progressive breakdown highlights GeoChain's utility as a diagnostic benchmark. By tracking model accuracy at each reasoning step, researchers can understand how performance degrades under deeper spatial inference chains.

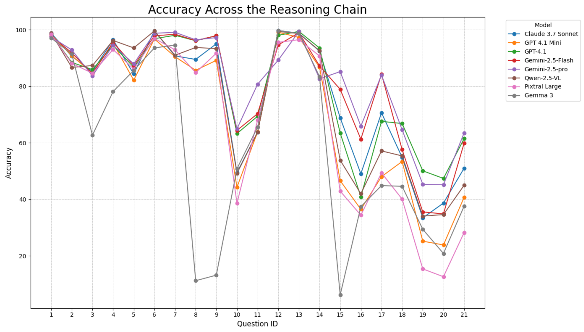


Figure 4: Average pass score across the 21-step Geochain reasoning chain

4.5 Fine-Tuning Study

Our primary analysis revealed a critical failure point in state-of-the-art MLLMs: their inability to consistently link visual cues to a final geographic location through sequential reasoning. This finding raised a pivotal question: is this a fundamental limitation of current models, or is it a tractable skill that can be taught?

To address this, we conducted a targeted fine-tuning experiment with a clear objective: to determine if training on the GeoChain framework could teach a model to forge this connection between visual evidence and final localization - the very skill our initial analysis found lacking - and then generalize this ability to entirely unseen cities. For this study, we fine-tuned the open-source **Qwen-VL** (9.6B) (Bai et al., 2023) model using Low-Rank Adaptation (LoRA) for 5 epochs on a modest set of 900 samples from five geographically diverse cities (Moscow, Berlin, Bangkok, Sao Paulo, and Phoenix). We then evaluated on a held-out set of 850 samples from five entirely different, unseen cities (Budapest, Ottawa, Helsinki, Melbourne, and Amsterdam).

The results of this brief intervention are striking. As shown in Table 4, fine-tuning boosted the **overall pass score by over 20 percentage points** (46.8% to 68.2%) and **nearly halved the mean localization error**, demonstrating a significant improvement in final prediction accuracy. Fine-tuning

Table 3: Pass score (%) by question category.

Model	Visual	Terrain	Geo Localization	Cultural	Exact Loc.
Claude 3.7 Sonnet	**92.8**	84.7	69.4	67.4	51.0
GPT-4.1 Mini	92.3	78.7	64.1	56.8	40.7
GPT-4.1	91.8	84.8	**76.9**	68.3	61.5
Gemini-2.5-Flash	92.4	86.0	73.5	75.3	59.8
Gemini-2.5-pro	92.1	**87.4**	76.8	**77.9**	**63.5**
Qwen 2.5 VL	92.3	81.6	70.8	62.0	45.0
Pixtral Large	91.0	78.7	61.1	53.9	28.2
Gemma 3	82.3	70.7	55.2	35.0	37.6

Table 4: Overall performance comparison before and after fine-tuning.

Model	Pass Score (%)	Mean Dist. (km)	<25 km (%)	<200 km (%)	<750 km (%)
Base	46.8	5866.7	0.15	1.0	18.7
Finetuned	**68.2**	**2963.4**	**6.3**	**15.5**	**40.5**

Table 5: City-level localization accuracy (<25 km) by image difficulty.

Model	Easy	Medium	Hard
Base	0.31%	0.00%	0.00%
Finetuned	**4.8%**	**7.2%**	**7.5%**

Table 6: Pass score (%) by question difficulty.

Model	Easy	Medium	Hard
Base	66.5%	40.5%	2.1%
Finetuned	**91.1%**	**63.0%**	**12.9%**

proved particularly effective at addressing the base model's key weaknesses:

- **Geospatial Grounding:** Localization accuracy on "Hard" images, a primary challenge for the base model, jumped from 0.00% to **7.5%** (Table 5), indicating that the model learned to connect visual evidence to specific geographic contexts.

- **Complex Reasoning:** The model's capacity for nuanced inference saw the most dramatic gains. The pass score on "Hard" questions increased more than **six-fold**, from 2.1% to **12.9%** (Table 6).

These findings, achieved with limited data, serve a crucial diagnostic purpose. The fact that targeted training on GeoChain so effectively remedies the specific deficiencies it identifies poor grounding and brittle multi-step inference validates its utility beyond mere evaluation. It establishes GeoChain as a fine-grained diagnostic framework that provides a clear and actionable roadmap for targeted model improvement.

5 Conclusion

This paper introduced GeoChain, a large-scale, chain-of-thought benchmark designed to dissect multimodal geographic reasoning in MLLMs using street-level imagery and a 21-step diagnostic framework. Our evaluations on the curated GeoChain Test-Mini subset reveal that even leading MLLMs exhibit significant deficiencies in visual grounding, reasoning consistency, and localization accuracy, particularly as task and visual complexity escalate. By enabling a granular, step-by-step analysis, GeoChain moves beyond simple end-task accuracy to pinpoint these critical failure modes, thereby providing an essential diagnostic resource and methodology. We anticipate that GeoChain will steer future research towards developing more robust, geographically aware, and reliable AI systems capable of nuanced real-world understanding.

6 Limitations

While GeoChain offers a novel diagnostic approach, we acknowledge several limitations. GeoChain is built upon the Mapillary Street-Level Sequences training split; consequently, while our chain-of-thought reasoning framework and the overall task are novel, there is a potential that MLLMs have encountered these specific visual scenes or highly similar ones during their extensive pre-training. Evaluating performance on truly "unseen" street-level imagery is an inherent challenge for the field, given the ubiquity of data from sources like Google Street View (Anguelov et al., 2010) and OpenStreetMap (Haklay and Weber, 2008), meaning that performance assessments may partly reflect familiarity with certain visual data rather than solely generalization to entirely new scenes. Additonally, the underlying geographical distribution of the images, though diverse, retains some skew, potentially affecting the generalizability of the findings in all urban contexts. Furthermore, our locatability score's precision is contingent upon the accuracy of an upstream semantic segmentation model, which could introduce noise into the difficulty stratification.

7 Usage of Generative AI tools

We utilized Generative AI tools to help improve the language, phrasing, and readability of this manuscript.

References

Dragomir Anguelov, Carole Dulong, Daniel Filip, Christian Frueh, Stéphane Lafon, Richard Lyon, Abhijit Ogale, Luc Vincent, and Josh Weaver. 2010. Google street view: Capturing the world at street level. *Computer*, 43(6):32–38.

Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5287–5297.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Ron Campos, Ashmal Vayani, Parth Parag Kulkarni, Rohit Gupta, Aritra Dutta, and Mubarak Shah. 2025. GAEA: A geolocation aware conversational model. *Preprint*, arXiv:2503.16423.

Yuchen Cao, Nilay Pande, Ayush Jain, Shikhar Sharma, Gabriel Sarch, Nikolaos Gkanatsios, Xian Zhou, and Katerina Fragkiadaki. Embodied symbiotic assistants that see, act, infer and chat.

Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. 2021. Per-pixel classification is not all you need for semantic segmentation. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, pages 17864–17875. Curran Associates, Inc.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

GeoGuessr. 2013. Geoguessr - let's explore the world! Accessed: [Date you accessed the website].

Google. 2025a. Gemini 2.5 flash is now in preview. https://blog.google/products/gemini/gemini-2-5-flash-preview/. Accessed 2025-05-18.

Google. 2025b. Gemini 2.5 pro model. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/gemini-2-5-thinking. Accessed through the Gemini API on [Date of access] or Vertex AI.

Muki Haklay and Patrick Weber. 2008. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18.

James Hays and Alexei A. Efros. 2008. IM2GPS: estimating geographic information from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE.

Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan David Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Andrew Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, Mehmet Gunturkun, Gabriel Huang, David Vazquez, Dava Newman, Yoshua Bengio, Stefano Ermon, and Xiao Xiang Zhu. 2023. Geo-bench: Toward foundation models for earth monitoring. *Preprint*, arXiv:2306.03831.

Ling Li, Yu Ye, Bingchuan Jiang, and Wei Zeng. 2024. Georeasoner: Geo-localization with reasoning in street views using a large vision-language model. In *International Conference on Machine Learning (ICML)*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Mistral AI. 2024. Introducing Pixtral Large. https://mistral.ai/news/pixtral-large. Accessed: 07/10/2025.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Nilay Pande and Suyash P. Awate. 2021. Generative deep-neural-network mixture modeling with semi-supervised minmax+em learning. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5666–5673.

Saksham Pramanik, Aayush Mundra, Ashutosh Mittal, Sreyas Mohan, Saim Wani, Shramay S Vernekar, Pranav M Dixit, Shanti Priya, Ankur Beniwal, Ojaswa Sharma, and Senthil Mani. 2024. Evaluating precise geolocation inference capabilities of vision language models. *Preprint*, arXiv:2502.14412.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

Zirui Song, Jingpu Yang, Yuan Huang, Jonathan Tonglet, Zeyu Zhang, Tao Cheng, Meng Fang, Iryna Gurevych, and Xiuying Chen. 2025. Geolocation with real human gameplay data: A large-scale dataset and human-like reasoning framework. *Preprint*, arXiv:2502.13759.

Claude 3.7 Sonnet. 2025. Claude 3.7 sonnet documentation. https://docs.anthropic.com/en/docs/overview. Anthropic AI Model.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Yicong Tian, Chen Chen, and Mubarak Shah. 2017. Cross-view image matching for geo-localization in urban environments. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1998–2006.

Carnegie Mellon University. 2023. Embodied symbiotic assistants that see, act, infer and chat.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Frederik Warburg, Søren Hauberg, Gregory D. D. Funke, and Yoko Yuki. 2020. Mapillary street-level sequences: A dataset for lifelong place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 931–940. IEEE.

Tobias Weyand, Ilya Kostrikov, and James Philbin. 2016. PlaNet - photo geolocation with convolutional neural networks. In *Computer Vision – ECCV 2016, Proceedings, Part II*, volume 9906 of *Lecture Notes in Computer Science*, pages 37–55. Springer.

Ruixiang Yang, Cheng Zhang, Lingxi Meng, He Wang, Xiaoyan Li, Yuke Li, Shuo Wang, Haoran Wei, Yiyang Li, Wentao Qu, Pengchuan Zhang, Jiazheng Xu, Bihan Wen, Diyi Yang, Kangkang Lu, Saurabh Gupta, Guanzhong Wang, Zhiqiang Shen, Baining Guo, and 3 others. 2024. VLMs as GeoGuessr masters—exceptional performance, hidden biases, and privacy risks. *Preprint*, arXiv:2502.11163.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 633–641. IEEE.

A Appendix

A.1 Implementation Details

A.1.1 Questions

This section details the complete 21-question sequence (Table 7) that forms the core of the GeoChain benchmark, designed to evaluate the step-by-step geographic reasoning capabilities of Multimodal Large Language Models (MLLMs). Each question in the sequence is characterized by its rank, designated difficulty level (Easy, Medium, or Hard), expected response format (Binary, Multi-class, or Free-text), and its primary Question Category (Visual Cues, Geographical localization, Culture/Infrastructure, Terrain/Environment, or Exact Location). This comprehensive listing provides a transparent foundation for understanding the specific tasks underpinning the performance evaluations discussed throughout this paper.

A.1.2 System Prompt

To guide the Multimodal Large Language Models (MLLMs) and standardize their responses for the GeoChain benchmark tasks, the following system prompt was consistently employed:

> **System Prompt**
>
> You are an accurate geolocation model. Given the image, answer the following questions in order. Please provide your best guess. Each question is also provided with question type. For Binary questions, answer Yes/No only. For Multiclass questions, answer as one of the provided options in brackets. Final question type is a free text question, answer it as a free string text. If you are not sure about the answer, give your best guess. Answer format should be a json dict with question indices as keys (0 indexed) and values as Answer: <answer>, Reasoning: <reasoning>.

A.1.3 Tools and Infrastructure

The execution of model inference was managed by Promptfoo[1], a platform that ensures reproducibility in benchmarking by offering versatile prompt configuration and effective API linkage. We used the transformers library in Hugging Face[2]; to run the MaskFormer model for computing segmentation masks. These calculations were performed on an NVIDIA GeForce RTX 3060 graphics processing unit.

A.2 Haversine Distance

Haversine Distance the shortest distance over the Earth's surface between the predicted and ground-truth coordinates, assuming a spherical Earth.

The Haversine formula is given by:

$$\Delta\phi = \phi_2 - \phi_1$$

$$\Delta\lambda = \lambda_2 - \lambda_1$$

$$a = \sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\Delta\lambda}{2}\right)$$

$$d = 2R \cdot \arcsin\left(\sqrt{a}\right)$$

Here, d is the Haversine distance between two points (ϕ_1, λ_1) and (ϕ_2, λ_2). This metric provides an interpretable and robust way to measure geographic prediction error.

A.3 Additional Analysis

A.3.1 Image Difficulty vs Question Difficulty Interaction

To analyze how visual and reasoning difficulty interact, we compute a two-dimensional pass rate matrix over **question difficulty** (Easy, Medium, Hard) and **image difficulty** (Easy, Medium, Hard). Table 8 presents this breakdown for each model.

We observe a consistent trend across all models: accuracy declines with both increasing *image* difficulty and *question* difficulty. Importantly, hard questions on hard images represent the most challenging setting, with pass rates often below 40%—even for state-of-the-art models.

Gemini-2.5-pro shows the strongest resilience across the board, maintaining high scores even on hard questions in ambiguous scenes. In contrast, Claude 3.7 Sonnet and GPT 4.1 Mini exhibit large drops in performance under compounding difficulty, confirming their brittleness in multi-factor reasoning.

This matrix allows us to quantify model *sensitivity to visual ambiguity* and pinpoint failure modes. For example, a model that performs well on hard questions from easy images but poorly on the same questions from hard images may lack robustness in interpreting noisy visual cues. Conversely, a model that fails uniformly on hard questions indicates weaknesses in logical chaining or symbolic inference. Together, this analysis emphasizes the need for benchmarks that probe cross-modal interactions, rather than evaluating visual or linguistic difficulty in isolation.

A.3.2 Breakdown by Question Difficulty

To better understand how models handle increasing reasoning complexity, we group questions by their annotated difficulty levels: **Easy**, **Medium**, and **Hard**. These difficulty tags were assigned manually based on the subtlety, required external knowledge, and ambiguity of each question.

Across all models, accuracy decreases consistently with question difficulty. Gemini-2.5-pro achieves the highest pass rates at all levels, followed closely by Gemini-2.5-Flash and GPT-4.1. Interestingly, Claude 3.7 Sonnet and GPT 4.1 Mini both exhibit sharp drops on hard questions, with performance falling below 45% and 35%, respectively.

These findings suggest that while many models can answer surface-level geographic questions

[1] https://www.promptfoo.dev

[2] https://huggingface.co/docs/transformers/en/index

Table 7: The GeoChain 21-Step Benchmark Question Set.

Rank	Difficulty	Question	Question Type	Question Category
1	Easy	Do you see any boats or ships?	Binary	Visual Cues
2	Easy	Do you see one or more of the following vehicles: Bus, Truck, Car, Van, Motorbike, Minibike, Bicycle?	Binary	Visual Cues
3	Easy	Can you see any traffic lights?	Binary	Visual Cues
4	Easy	Can you see any flag?	Binary	Visual Cues
5	Easy	Would you say this location is near the Equator?	Binary	Geographical localization
6	Easy	Does this location seem to be close to the Poles?	Binary	Geographical localization
7	Easy	Is this place located in the Northern Hemisphere?	Binary	Geographical localization
8	Easy	Which continent best describes where this location is? (7 continents: North America/South America/Europe/Africa/Asia/Oceania/Antarctica)	Multiclass	Geographical localization
9	Medium	What side of the road do vehicles drive on here? (Left/Right)	Multiclass	Culture/Infrastructure
10	Medium	What country is this place located in?	Free-text	Geographical localization
11	Medium	Is this place near coast?	Binary	Terrain/Environment
12	Medium	Does this location appear to be an island?	Binary	Terrain/Environment
13	Easy	Is this place located in a desert region?	Binary	Terrain/Environment
14	Easy	Does this location seem to be in a mountainous or hilly region?	Binary	Terrain/Environment
15	Medium	What is the most likely climate type for this location? (5 main climate types: Tropical/Dry/Temperate/Continental/Polar)	Multiclass	Terrain/Environment
16	Easy	Does this place look like a big city?	Binary	Culture/Infrastructure
17	Medium	Would you classify this place as a small town?	Binary	Culture/Infrastructure
18	Hard	What language(s) are most likely spoken at this place?	Free-text	Culture/Infrastructure
19	Hard	Can you name the state or province this place belongs to?	Free-text	Geographical localization
20	Hard	What is the name of the city, town, or village seen here?	Free-text	Geographical localization
21	Hard	Based on everything observed, what are the latitude and longitude coordinates of this place? Please give a tuple of float coordinates (lat, lon)	Free-text	Exact Location

Table 8: Pass score (%) across question difficulty and image difficulty. Each row shows performance on a given question difficulty across images of increasing ambiguity.

Model	Question Difficulty	Easy Images	Medium Images	Hard Images
Claude 3.7 Sonnet	Easy	89.3	89.1	87.7
	Medium	76.0	75.7	72.0
	Hard	45.8	52.7	34.8
GPT 4.1 Mini	Easy	86.7	86.7	84.2
	Medium	66.1	67.3	62.1
	Hard	37.4	44.6	27.9
GPT-4.1	Easy	87.9	87.5	84.3
	Medium	75.5	76.5	71.8
	Hard	51.8	60.0	43.3
Gemini-2.5-Flash	Easy	90.7	91.0	89.4
	Medium	76.7	77.8	74.2
	Hard	47.3	54.9	38.4
Gemini-2.5-pro	Easy	**91.6**	**91.3**	**89.8**
	Medium	**78.2**	**79.9**	**75.7**
	Hard	**52.4**	**61.6**	**45.9**

Table 9: Pass score (%) by question difficulty.

Model	Easy	Medium	Hard
Claude 3.7 Sonnet	88.7	74.6	44.5
GPT 4.1 Mini	85.9	65.2	33.4
GPT-4.1	87.3	75.8	54.7
Gemini-2.5-Flash	90.8	76.2	51.3
Gemini-2.5-pro	**91.1**	**78.4**	**55.1**

accurately, their reasoning falters as complexity increases especially when fine-grained localization or symbolic inference is required. The relatively better performance of Gemini-2.5-pro on hard questions indicates more stable multi-hop reasoning or greater robustness to subtle visual signals.

A.3.3 Breakdown by Question Type

To assess how models handle varying degrees of response constraint, we analyzed Pass Scores across three fundamental question types: Binary, Multiclass, and Free-text, with results presented in Table 10 and Figure 6. This breakdown reveals a distinct performance hierarchy directly correlated with the open-endedness of the required answer.

Across all evaluated MLLMs, a clear difficulty gradient was observed: Binary questions yielded the highest success rates, followed by Multiclass questions, with Free-text questions proving to be the most challenging by a substantial margin. For instance, Gemini-2.5-pro achieved 88.9% on Binary and an exceptional 92.9% on Multiclass questions, but its score dropped to 56.7% for Free-text tasks. This pattern of significantly lower performance on Free-text questions was universal, underscoring the inherent difficulty in precise, open-ended generation and factual recall compared to selecting from constrained options.

In the structured formats, Gemini-2.5-pro consistently led, achieving the top scores for both Binary (88.9%) and Multiclass (92.9%) questions, with Gemini-2.5-Flash also performing strongly. Notably, for the more demanding Free-text questions, GPT-4.1 emerged as the top performer with a Pass Score of 57.8%, slightly ahead of Gemini-2.5-pro (56.7%). This suggests a particular strength in GPT-4.1's generative capabilities for unconstrained answers. Claude 3.7 Sonnet demonstrated robust performance on Binary (86.2%) and Multiclass (84.5%) questions, often comparable to GPT-4.1, but its accuracy significantly declined on Free-text questions (45.5%), reaffirming its challenges with precise, unprompted generation. As anticipated, GPT-4.1 Mini generally recorded the lowest scores across all types. This analysis by question type effectively highlights that while current MLLMs are largely proficient with constrained-choice tasks, open-ended free-text responses remain a key area

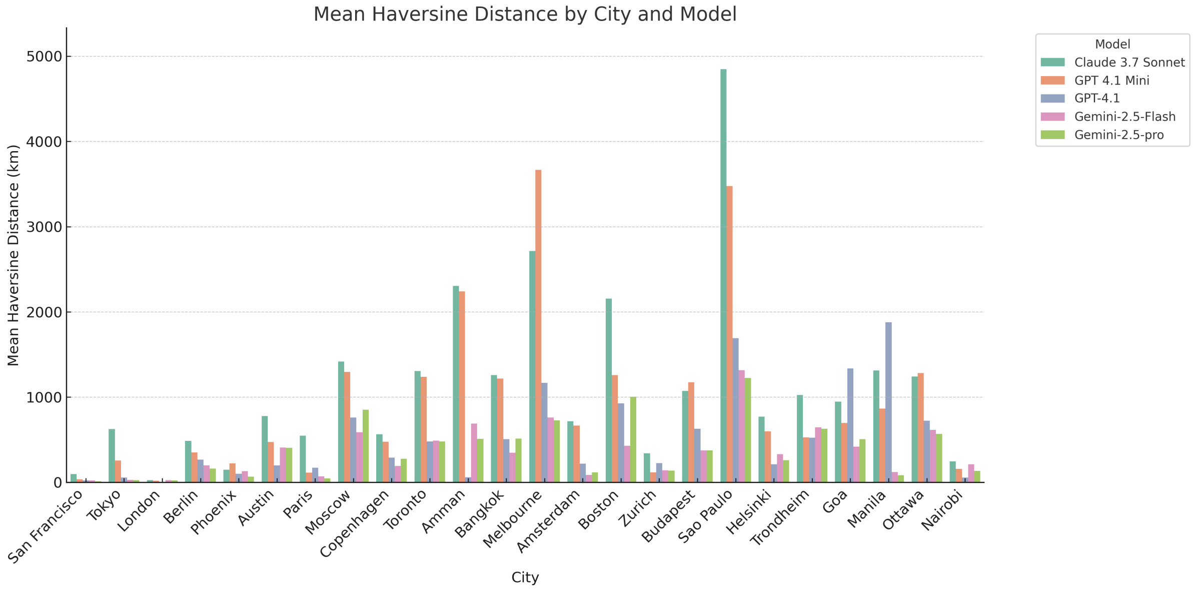


Figure 5: Mean Haversine distance (km) by city and model. Larger values indicate poor localization precision.

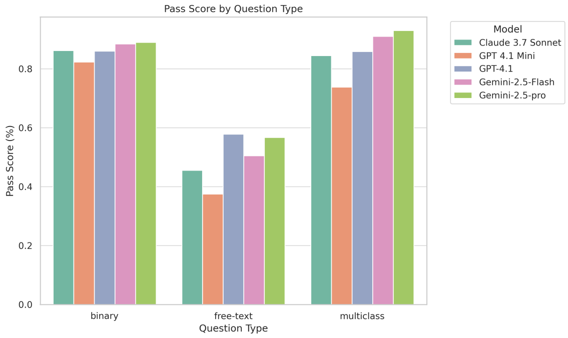


Figure 6: Model vs Question Type

for improvement.

Table 10: Pass score (%) by question type.

Model	**Binary**	**Multiclass**	**Free-text**
Claude 3.7 Sonnet	86.2	84.5	45.5
GPT-4.1 Mini	82.3	73.8	37.5
GPT-4.1	85.9	85.8	57.8
Gemini-2.5-Flash	88.5	90.9	50.5
Gemini-2.5-pro	**88.9**	**92.9**	**56.7**

A.4 Breakdown by City

A city-level view (Fig. 7) shows that performance is far from uniform:

Gemini-2.5-pro is the most stable, topping the leaderboard in 20 / 24 cities and exceeding 85% accuracy in visually distinctive urban centres such as Tokyo, Zurich and Toronto. Gemini-2.5-Flash and GPT-4.1 follow closely, maintaining more than **75%** accuracy in most regions. Performance on Claude 3.7 Sonnet and GPT 4.1 Mini fluctuate sharply: they perform competitively in cue-rich European cities (Paris, Berlin) but collapse in visually ambiguous locales (Nairobi, São Paulo, Amman). Mean Haversine error (Fig. 5) confirms the pattern: Gemini-2.5-pro keeps errors below 300 km in nearly every city, whereas Claude and GPT 4.1 Mini exceed 1000 km in several cases (Helsinki, Melbourne, São Paulo).

These results highlight how regional factors such as vegetation, signage language, traffic orientation and architectural style strongly modulate geolocation accuracy.

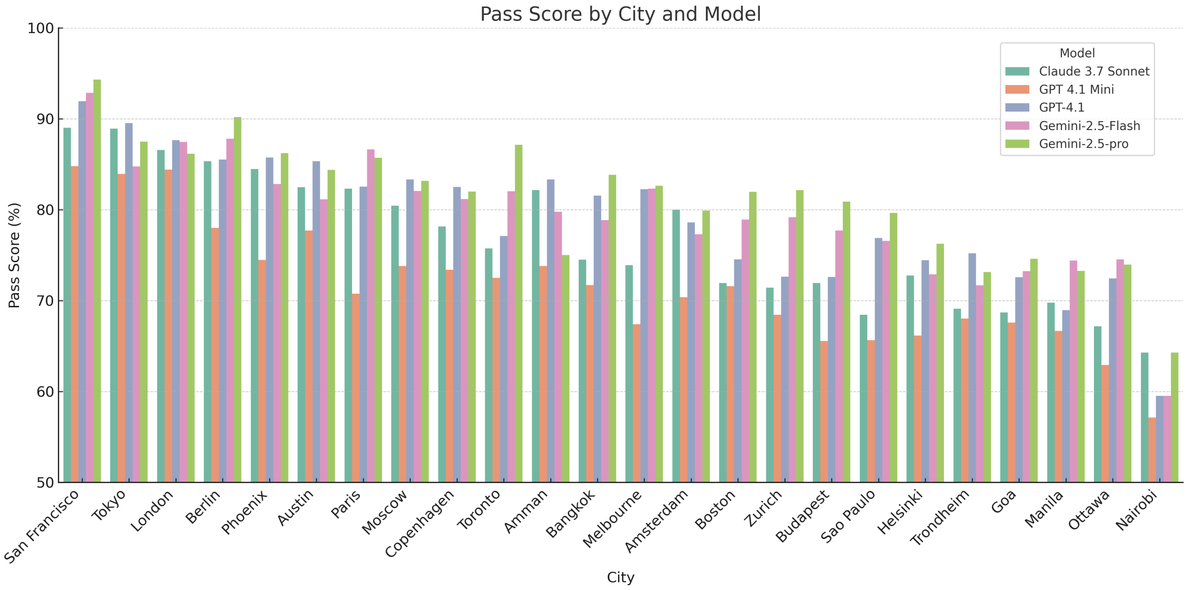


Figure 7: Pass score (%) by city, highlighting the influence of geographical location on model accuracy.