GeoChain: Multimodal Chain-of-Thought for Geographic Reasoning

Sahiti Yerramilli [*1] **Nilay Pande** [*2] **Rynaa Grover** [*1] **Jayant Sravan Tamarapalli** [*1]

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 contemporary MLLMs (GPT-4.1 variants, Claude 3.7, Gemini 2.5 variants) 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: https://github.com/sahitiy/geochain

Dataset: https://huggingface.co/datasets/sahitiy51/geochain

1. Introduction

Recent advancements in large vision-language models (VLMs) increasingly viewed as foundational components of sophisticated world models underscore their growing proficiency in general visual understanding (Google, 2025a; OpenAI et al., 2024; Wang et al., 2024; Dai et al., 2023; Liu et al., 2023). However, the extent to which these models genuinely understand complex domains in the real world like geography, particularly their capacity for nuanced, structured geographic reasoning, remains significantly under-investigated. Such reasoning, which involves inferring location by synthesizing visual cues with spatial and cultural knowledge, is not only vital for diverse applications in the real world but also serves as a critical domain for assessing the groundedness and fidelity of these models' internal world representations. Current evaluation techniques often fall short, with most benchmarks emphasizing end-task prediction accuracy while neglecting to probe the step-by-step inferential processes that signify deeper understanding. GeoChain directly addresses this methodological void. This novel multimodal benchmark leverages 1.46 million Mapillary street-level images (Warburg et al., 2020), each paired with a 21-step chain-of-thought (CoT) question sequence that guides models from broad (e.g., continent) to precise (e.g., city, coordinates) geolocation. This methodology yields over 30 million question-answer pairs, categorized by reasoning type (visual, spatial, cultural, precise geolocation) and annotated with difficulty, facilitating granular diagnostic insights. Figure 1 presents a visual instance of GeoChain's components.

For focused evaluation, we curated GeoChain Test-Mini, a challenging 2,088-image subset. Its creation involved adapting a 'locatability score', drawing from GeoReasoner (Li et al., 2024) and computed using features from a pretrained MaskFormer model (Cheng et al., 2021), enabling stratification into Easy, Medium, and Hard tiers. Our contributions are threefold: (1) The GeoChain benchmark framework itself, offering extensive, diagnostically rich data for step-by-step geographic reasoning evaluation. (2) Augmentation of rich semantic labels to the images and the development of GeoChain Test-Mini through a locatability score, providing a valuable community resource. (3) A comprehensive benchmarking of leading MLLMs on GeoChain Test-Mini, yielding detailed analyses of their geographic reasoning capabilities and failure modes.

2. Related Work

The field of image-based geolocation has evolved from early database matching techniques and deep learning-based coordinate prediction (Hays & Efros, 2008; Weyand et al.,

[*]Equal contribution [1]Google, Mountain View, CA, USA [2]Waymo, Mountain View, CA, USA. Correspondence to: Sahiti Yerramilli <sahitiy@google.com>, Nilay Pande <nilayp@waymo.com>.

ICML 2025 Workshop on Assessing World Models.

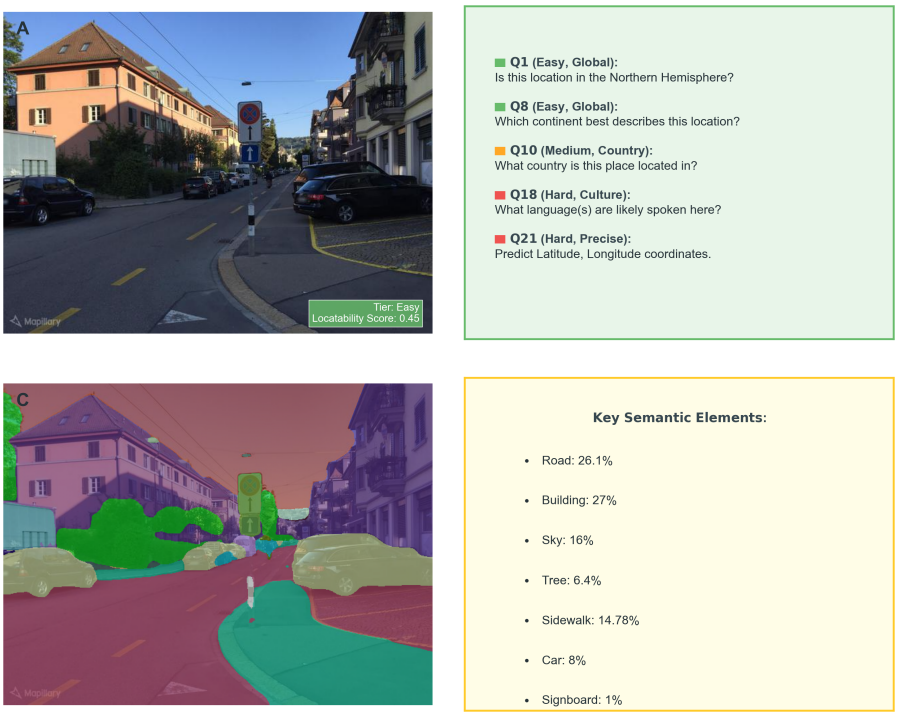

Figure 1. Components of a GeoChain instance: **(Top-Left)** Easy Mapillary Street-Level Sequences (MSLS) image with locatability score of 0.45. **(Top-Right)** Example chain-of-thought questions with difficulty indicators. **(Bottom-Left)** Derived semantic segmentation map. **(Bottom-Right)** Extracted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

2016; Arandjelovic et al., 2016; Tian et al., 2017) primarily targeting endpoint accuracy towards more sophisticated multimodal reasoning. Within this advanced landscape, several approaches leverage human-like inference from gameplay (Li et al., 2024; Song et al., 2025), focus on precise coordinate outputs (Pramanik et al., 2024; Yang et al., 2024), or utilize dynamic metadata for query generation (Campos et al., 2025). Other benchmarks address distinct domains such as remote sensing (Lacoste et al., 2023). GeoChain distinguishes itself by introducing a benchmark focused on structured, step-by-step diagnostic reasoning. It employs a fully static, image-grounded evaluation with a standardized 21-step chain-of-thought question sequence for each ground-level image. This design facilitates controlled and fair benchmarking of different models' inherent geographic reasoning capabilities spanning visual, spatial, and cultural cues without reliance on model fine-tuning, gameplay dynamics, or volatile external metadata, thereby offering unique insights into the consistency and granularity of the reasoning process itself.

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), captured under varied conditions, 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, each image is augmented with a semantic segmentation map detailing its composition across 150 classes spanning objects, environmental, and architectural features. These maps are generated using MaskFormer (Cheng et al., 2021) pre-trained on ADE20K (Zhou et al., 2017). We then calculate the percentage of image area covered by each category (e.g., 'sky' 30%, 'road' 15%), using these semantic coverage statistics to inform ground-truth

generation for many of the benchmark's visual questions.

3.2. Locatability Score Computation

To systematically evaluate model performance across varying visual ambiguity, we adopt the locatability scoring methodology from (Li et al., 2024). This score (0–1) reflects how visually distinctive a location is, higher scores denote easily identifiable scenes. Briefly, we compute scores by measuring semantic similarity between common cues used by expert GeoGuessr players and MaskFormer segmentation class labels. This similarity informs class weights, which are then aggregated based on their spatial coverage in each image. Using this score, we divide GeoChain images into three difficulty tiers: **Hard** if score $\in [0.12, 0.22)$, **Medium** if score $\in [0.22, 0.45)$, and **Easy** if score $\in [0.45, 0.6)$. Refer to Appendix A.2 for methodological details.

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.3.1. 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., 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 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. 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. We randomly sample unique image sequences across all available cities within each locatability tier. These 2100 candidates underwent manual visual inspection, where 12 images with critical quality issues 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, 2025b) and Gemini 2.5 Pro (Google,

2025a) 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

The accuracy of predicted geographic coordinates (Question 21) is evaluated using the **Haversine distance**, which calculates the shortest distance over the Earth's surface between the predicted and ground-truth locations (details in Section A.4). Overall performance is measured by the **Pass Score**, representing the average fraction of correctly answered questions across the 21-step reasoning chain. For most questions, correctness is determined by matching the ground-truth answer based on its type (Binary, MultiClass, Free-Text). Critically, the final coordinate prediction contributes to the Pass Score as correct if its Haversine distance from the true location is less than 50km.

4.2. Overall Model Performance

MLLM performance on GeoChain Test-Mini (Table 1) reveals specialized capabilities. Gemini-2.5-pro excels in multi-step reasoning (81.84% pass score), while Gemini-2.5-Flash demonstrates superior localization precision (445.24 km mean error), suggesting distinct optimization for these skills. This divergence underscores that broad inferential ability and precise geolocalization are distinct skills. Although GPT-4.1 is competitive, notable localization inaccuracies from Claude 3.7 Sonnet (1289.04 km) and GPT-4.1 Mini (1194.77 km) highlight robust geospatial grounding as a primary MLLM developmental hurdle.

Threshold-based localization metrics (City $<$25km, Region $<$200km, Country $<$750km) further differentiate models. Gemini-2.5-pro leads in City-level precision (59.38%), while Gemini-2.5-Flash excels at Region (70.02%) and Country (90.31%) granularities. GPT-4.1 also achieves strong City-level accuracy (57.84%). In contrast, Claude 3.7 Sonnet (40.34% City) and GPT-4.1 Mini (48.61% City) struggle significantly at these finer scales. These granular metrics effectively highlight that achieving reliable, high-confidence City-level precision is a key challenge across MLLMs.

4.3. Breakdown by Image Difficulty

Analyzing model performance by image difficulty (Table 3) 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 2), 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 score, 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.

5. Conclusion

We introduced GeoChain, a large-scale, 21-step chain-of-thought benchmark using street-level imagery to diagnose MLLM geographic reasoning. Evaluations on GeoChain Test-Mini subset show leading MLLMs struggle with visual grounding, reasoning consistency, and localization, especially with increased complexity. GeoChain's step-by-step analysis pinpoints these failures beyond end-task accuracy, offering a key diagnostic resource to foster more robust, geographically aware AI. A more detailed exploration of further

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

Table 2. 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**

Table 3. Performance by image difficulty. Pass Score (%) 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**

analytical dimensions is provided in A.5. A discussion of GeoChain's limitations is also presented in A.6.

References

Anguelov, D., Dulong, C., Filip, D., Frueh, C., Lafon, S., Lyon, R., Ogale, A., Vincent, L., and Weaver, J. Google street view: Capturing the world at street level. *Computer*, 43(6):32–38, 2010. doi: 10.1109/MC.2010.170.

Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., and Sivic, J. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5287–5297, 2016.

Campos, R., Vayani, A., Kulkarni, P. P., Gupta, R., Dutta, A., and Shah, M. GAEA: A geolocation aware conversational model, 2025.

Cheng, B., Schwing, A. G., and Kirillov, A. Per-pixel classification is not all you need for semantic segmentation. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, pp. 17864–17875. Curran Associates, Inc., 2021.

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. URL https://arxiv.org/abs/2305.06500.

GeoGuessr. Geoguessr - let's explore the world!, 2013. URL https://www.geoguessr.com/. Accessed: [Date you accessed the website].

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

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

Haklay, M. and Weber, P. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18, 2008. doi: 10.1109/MPRV.2008.80.

Hays, J. and Efros, A. A. IM2GPS: estimating geographic information from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8. IEEE, June 2008. doi: 10.1109/CVPR.2008.4587784.

Lacoste, A., Lehmann, N., Rodriguez, P., Sherwin, E. D., Kerner, H., Lütjens, B., Irvin, J. A., Dao, D., Alemohammad, H., Drouin, A., Gunturkun, M., Huang, G., Vazquez, D., Newman, D., Bengio, Y., Ermon, S., and Zhu, X. X. Geo-bench: Toward foundation models for earth monitoring, 2023. URL https://arxiv.org/abs/2306.03831.

Li, L., Ye, Y., Jiang, B., and Zeng, W. Georeasoner: Geolocalization with reasoning in street views using a large vision-language model. In *International Conference on Machine Learning (ICML)*, 2024.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X., and Wen, J.-R. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=xozJw0kZXF.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning, 2023.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Pramanik, S., Mundra, A., Mittal, A., Mohan, S., Wani, S., Vernekar, S. S., Dixit, P. M., Priya, S., Beniwal, A., Sharma, O., and Mani, S. Evaluating precise geolocation inference capabilities of vision language models, 2024.

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., and Saenko, K. Object hallucination in image captioning. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J. (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1437. URL https://aclanthology.org/D18-1437/.

Song, Z., Yang, J., Huang, Y., Tonglet, J., Zhang, Z., Cheng, T., Fang, M., Gurevych, I., and Chen, X. Geolocation

with real human gameplay data: A large-scale dataset and human-like reasoning framework, 2025. URL https://arxiv.org/abs/2502.13759.

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

Tian, Y., Chen, C., and Shah, M. Cross-view image matching for geo-localization in urban environments. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1998–2006, 2017. doi: 10.1109/CVPR.2017.216.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., and Lin, J. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. URL https://arxiv.org/abs/2409.12191.

Warburg, F., Hauberg, S., Funke, G. D. D., and Yuki, Y. 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*, pp. 931–940. IEEE, June 2020.

Weyand, T., Kostrikov, I., and Philbin, J. PlaNet - photo geolocation with convolutional neural networks. In Leibe, B., Matas, J., Sebe, N., and Welling, M. (eds.), *Computer Vision – ECCV 2016, Proceedings, Part II*, volume 9906 of *Lecture Notes in Computer Science*, pp. 37–55. Springer, 2016. doi: 10.1007/978-3-319-46475-6_3.

Yang, R., Zhang, C., Meng, L., Wang, H., Li, X., Li, Y., Wang, S., Wei, H., Li, Y., Qu, W., Zhang, P., Xu, J., Wen, B., Yang, D., Lu, K., Gupta, S., Wang, G., Shen, Z., Guo, B., Jin, R., Zhu, S.-C., and Lu, H. VLMs as GeoGuessr masters—exceptional performance, hidden biases, and privacy risks, 2024.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., and Torralba, A. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 633–641. IEEE, July 2017.

A. Appendix

A.1. MSLS City-Wise Distribution

Figure 2 details the count of images per city, thereby illustrating the urban distribution within the MSLS dataset.

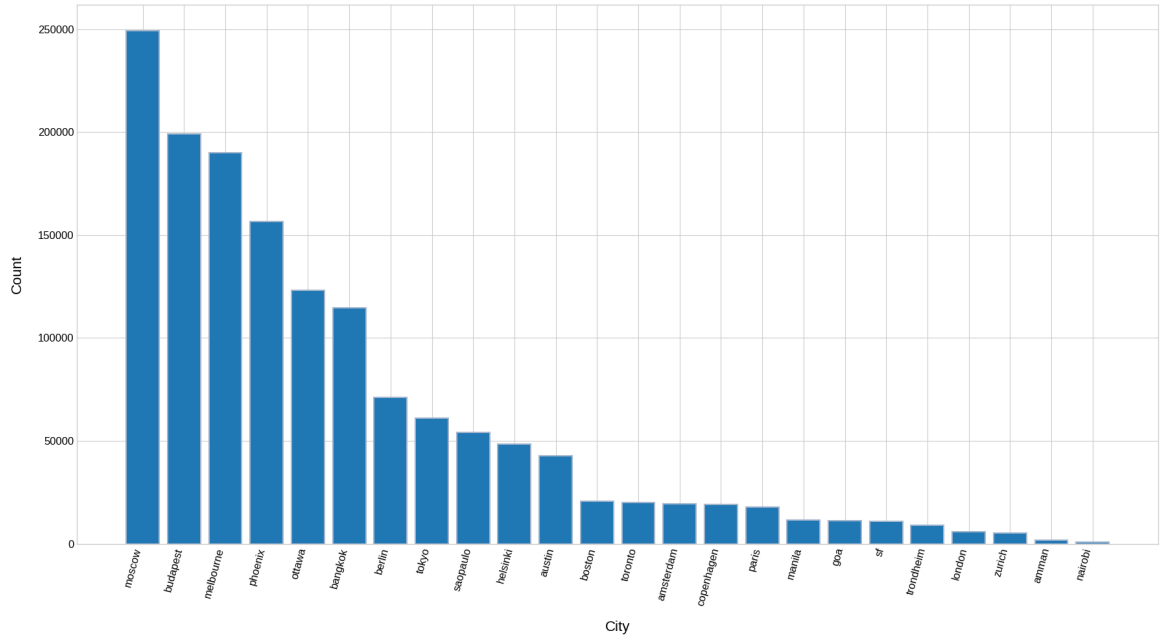

Figure 2. Count of images per city, illustrating the city distribution within the GeoChain dataset.

A.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 cues that humans, particularly proficient GeoGuessr players (GeoGuessr, 2013), rely on for geolocalization. The process involves several steps:

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: Easy, Medium,

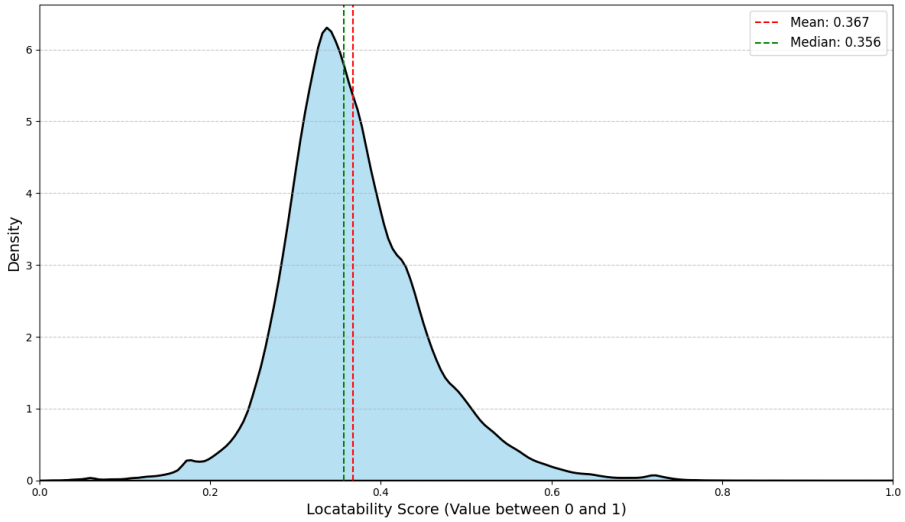

Figure 3. Distribution of Locatability Scores in GeoChain

and Hard. 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.

A.3. Implementation Details

A.3.1. Questions

This section details the complete 21-question sequence (Table 4) 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, Multiclass, 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.3.2. 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.3.3. 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.

[1] https://www.promptfoo.dev
[2] https://huggingface.co/docs/transformers/en/index

Table 4. 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

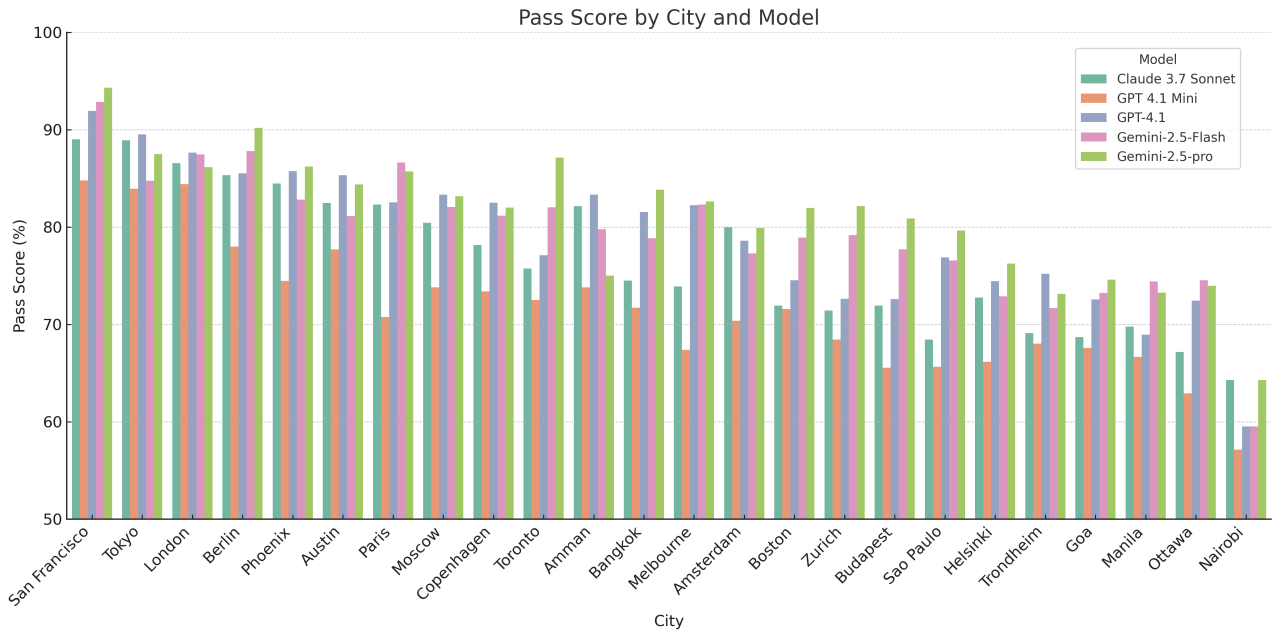

Figure 4. Pass score (%) by city, highlighting the influence of geographical location on model accuracy.

A.4. 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.5. Additional Analysis

A.5.1. Breakdown by City

A city-level view (Fig. 4) 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. 6) 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.

Table 5. 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**

A.5.2. 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 5 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.5.3. 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 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.

Table 6. 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**

A.5.4. Accuracy vs. Reasoning Depth

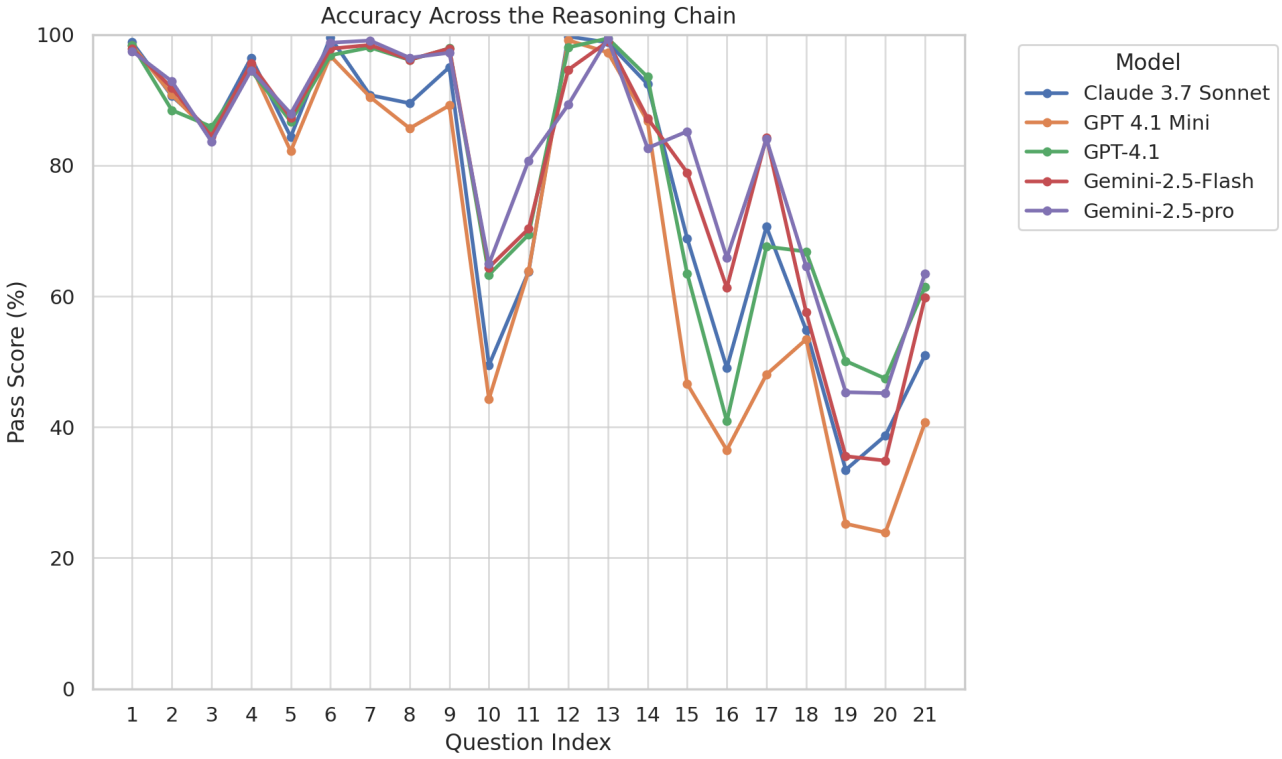

Figure 5. Average pass score across the 21-step Geochain reasoning chain. Accuracy decreases as questions progress from coarse global inference to fine-grained localization.

Figure 5 reveals a typical degradation pattern: All models perform well in the initial questions (1–9), which ask about visual or global cues such as vehicles, 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.

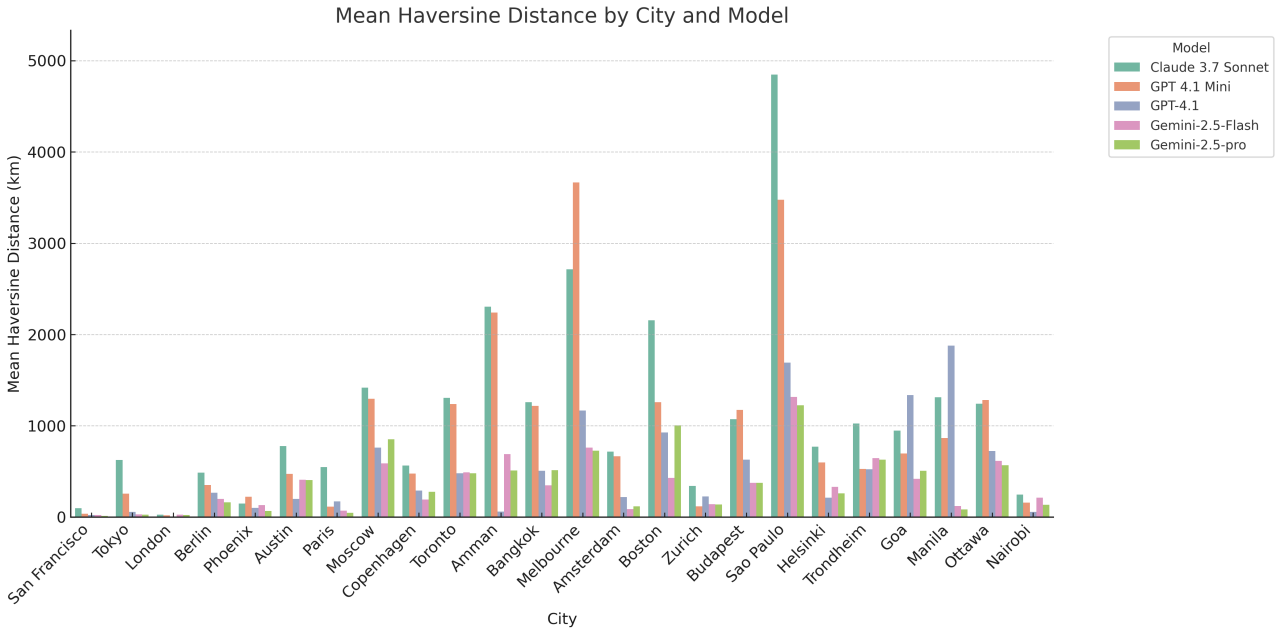

Figure 6. Mean Haversine distance (km) by city and model. Larger values indicate poor localization precision.

This progressive breakdown highlights GeoChain's utility as a diagnostic benchmark. By tracking model accuracy at each reasoning step, researchers can isolate failure modes (e.g. visual hallucination vs. failure to capture cultural cues) and understand how performance degrades under deeper spatial inference chains.

A.5.5. 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 7 and Figure 7. 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 for improvement.

A.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

Table 7. 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**

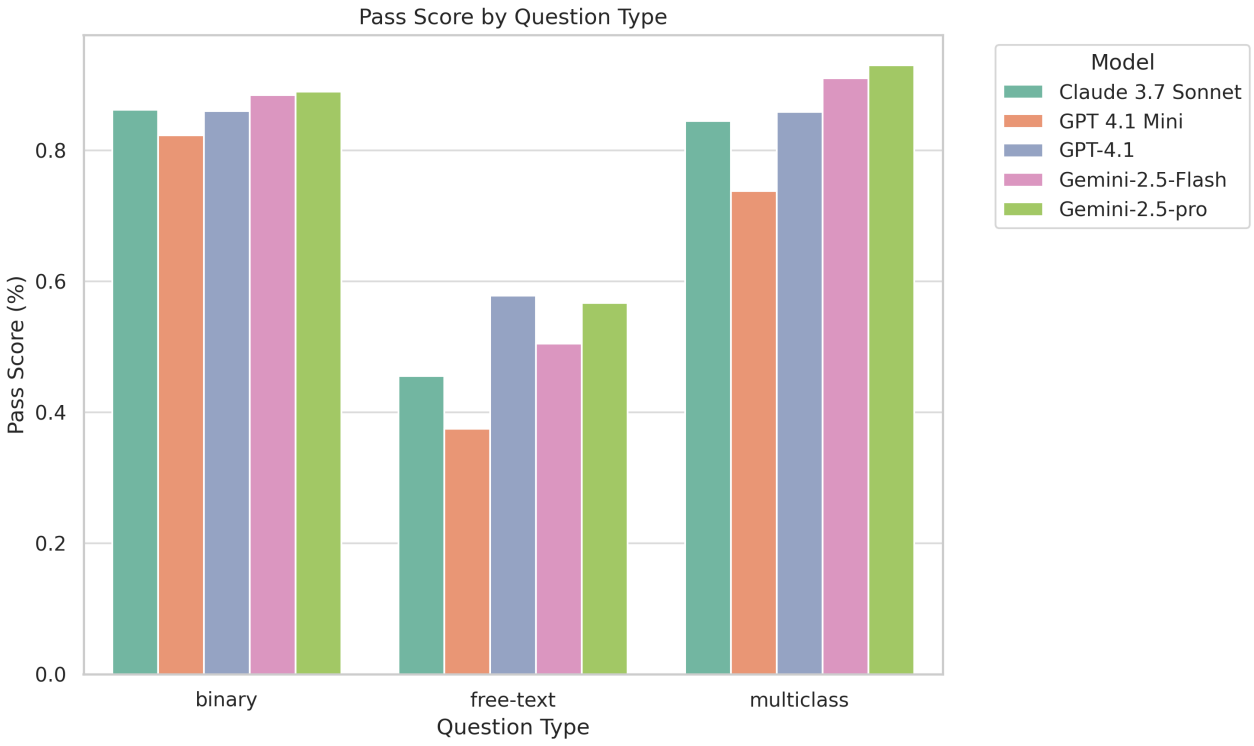

Figure 7. Model vs Question Type

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 & 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.