# Assessing the Geolocation Capabilities, Limitations and Societal Risks of Generative Vision-Language Models

**Oliver Grainge[1*], Sania Waheed[1*], Jack Stilgoe[2], Michael Milford[3], Shoaib Ehsan[1, 4]**

[1]University of Southampton,
[2]University College London,
[3]Queensland University of Technology,
[4]University of Essex

oeg1n18@soton.ac.uk s.waheed@soton.ac.uk j.stilgoe@ucl.ac.uk michael.milford@qut.edu.au s.ehsan@soton.ac.uk

## Abstract

Geo-localization is the task of identifying the location of an image using visual cues alone. It has beneficial applications, such as improving disaster response, enhancing navigation, and geography education. Recently, Vision-Language Models (VLMs) are increasingly demonstrating capabilities as accurate image geo-locators. This brings significant privacy risks, including those related to stalking and surveillance, considering the widespread uses of AI models and sharing of photos on social media. The precision of these models is likely to improve in the future. Despite these risks, there is little work on systematically evaluating the geolocation precision of Generative VLMs, their limits and potential for unintended inferences. To bridge this gap, we conduct a comprehensive assessment of the geolocation capabilities of 25 state-of-the-art VLMs on four benchmark image datasets captured in diverse environments. Our results offer insight into the internal reasoning of VLMs and highlight their strengths, limitations, and potential societal risks. Our findings indicate that current VLMs perform poorly on generic street-level images yet achieve notably high accuracy (61%) on images resembling social media content, raising significant and urgent privacy concerns.

## Introduction

Geo-localization, the task of identifying the geographic location of an image using only its visual content, is an important capability for applications such as disaster response, navigation, and geographic education. Traditional approaches have primarily relied on supervised learning with large geotagged datasets or on large-scale image retrieval methods. Recent advances in Vision-Language Models (VLMs) have revealed that these models possess emergent geo-localization abilities, even without task-specific training. Early work in this area focused on models such as CLIP (Vivanco, Nayak, and Shah 2023; Haas, Alberti, and Skreta 2023; Wu and Huang 2022), but more recent efforts have applied generative VLMs to this task (Zhou et al. 2024; Mendes et al. 2024; Waheed et al. 2025). These models have shown surprisingly strong performance on geo-localization benchmarks, often matching or exceeding the performance of specialized geo-localization models (Mendes et al. 2024).

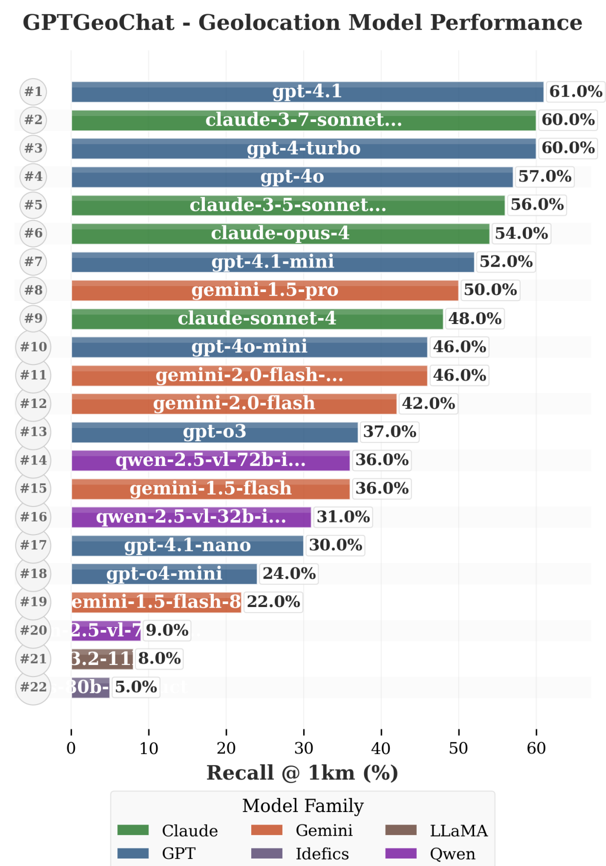


Figure 1: Leaderboard of 22 open-source, open-weight and closed-source vision–language models (VLMs) on the GPT-GeoChat dataset (Mendes et al. 2024). Models are ranked by Recall@1km, which is the percentage of predictions within 1 km of the ground truth. Models with 0% recall or a failure rate above 90% are excluded.

The rapid improvement of VLM-based geo-localization creates both major opportunities and serious risks. Their

*These authors contributed equally.

widespread availability and multipurpose nature makes regulation difficult. Precise localization from public images can be exploited for stalking, surveillance, or discrimination. As performance and accessibility improve, developers and users will look for new opportunities to generate value from geolocation. Without strong regulation, risks will escalate. We see risk in sociotechnical terms (Lazar and Nelson 2023): risk and safety are not just functions of the capabilities of AI models, but of the societal context in which these models sit, and in which they are developed, used and regulated. From this perspective, privacy extends beyond individual rights, serving as an important check on the power of those developing and deploying technologies with surveillance potential (Solove 2025). Rigorous evaluation of the capabilities and limitations of these models is therefore essential for responsible deployment and to inform policymakers about misuse and necessary safeguards.

Despite these concerns, there is limited systematic research on the geo-localization performance of generative VLMs. While prior work has examined data leakage and memorization in large language models (LLMs) (Carlini et al. 2020, 2022), few studies have investigated how generative VLMs perform in geo-localization across diverse, real-world contexts (Waheed et al. 2025).

To bridge this gap, we present a comprehensive black-box evaluation of 25 state-of-the-art generative VLMs, including open-source, open-weight and closed-source systems. We evaluate these models on four benchmark datasets that collectively capture a broad range of scenarios, including globally distributed images (im2gps (Hays and Efros 2008) and im2gps3k (Vo, Jacobs, and Hays 2017)), urban street-level imagery (OSV5 (Astruc et al. 2024)), and data curated to resemble images commonly posted on social media (GPT-GeoChat (Mendes et al. 2024)). In addition to measuring accuracy, we also account for cases where models fail to respond.

Our results provide the first large-scale comparative study of generative VLMs in geo-localization and highlight both their performance potential and the privacy risks that arise from their growing deployment in real-world applications. Our key contributions are:

- We benchmark 25 state-of-the-art generative VLMs on four diverse datasets, offering a detailed comparison of their strengths and limitations.
- We analyze the broader societal and ethical risks associated with VLM-based geo-localization, offering insights for future research and policy.

The remainder of this paper is organized as follows: Related Work summarizes prior research; Experimental Setup describes models, datasets, metrics, and methodology; Results and Analysis presents findings and comprehensive performance analysis; Discussion and Implications considers societal impact and broader implications; and Conclusion summarizes the study and future directions

## Related Work

This section covers prior research relevant to our work, covering both technical approaches to geo-localization and emerging concerns around privacy and societal implications.

### Geo-localization and Vision-Language Models

Traditional geo-localization research (Hays and Efros 2008) treats localization as a nearest-neighbor problem over a large-scale geotagged image database. This approach works well for landmark recognition (Arandjelovic et al. 2016; Weyand et al. 2020) but faces scalability challenges and perceptual aliasing at global levels (Muller-Budack, Pustu-Iren, and Ewerth 2018). To overcome these limitations, classification-based methods (Weyand, Kostrikov, and Philbin 2016; Muller-Budack, Pustu-Iren, and Ewerth 2018; Kordopatis-Zilos et al. 2021) partition the Earth into discrete geo-cells, enabling broader global reasoning but introduce trade-offs between spatial resolution and class imbalance (Kordopatis-Zilos et al. 2021; Haas et al. 2024).

Recent advances in Vision-Language Models (VLMs) have introduced a new paradigm for geo-localization. Models like CLIP (Radford et al. 2021) add semantic reasoning and contextual understanding to feature representations (Wu and Huang 2022; Haas et al. 2024; Luo et al. 2022; Zhou et al. 2024). More recently, generative VLMs such as GPT-4v (Achiam et al. 2023a) can infer geographic locations directly from images, in some cases surpassing specialized geo-localization models (Mendes et al. 2024; Waheed et al. 2025). Despite these advances, evaluations remain limited, with most studies testing only a few models and overlooking factors such as family, scale, training paradigm, and failure cases. This highlights the need for systematic, large-scale benchmarking of generative VLMs for geo-localization.

### Privacy and Societal Risks

The sensitivity of location data has long been a central concern in privacy research, as even seemingly innocuous information can reveal sensitive personal details. The *privacy paradox* highlights a persistent disconnect between individuals' stated concerns and their behavior: while people claim to value privacy, they often disclose personal data freely and fail to take protective measures (Gerber, Gerber, and Volkamer 2018; Hargittai and Marwick 2016; Kokolakis 2017). This gap between awareness and action is particularly dangerous for location data, which can reveal intimate aspects of daily life, including routines, habits, and social connections.

Prior studies have shown that even limited location traces can be used to reconstruct or predict a person's movements with surprising accuracy (Krumm 2022). While earlier research primarily focused on structured data sources, such as GPS logs or check-ins (Marmasse and Schmandt 2002; Ashbrook and Starner 2003; Chen et al. 2017), advances in VLMs have introduced new risks. Visual content shared online, such as social or travel photos, can now be reverse-engineered to reveal precise geographic locations (Mendes et al. 2024). This capability raises risks not only for individual privacy but also for public safety by enabling stalking, targeted harassment, or even large-scale surveillance.

| Model | R@1km (%) GPTGeoChat | Im2GPS | Im2GPS3k | R@100km (%) OSV5 | Avg. across datasets City Acc (%) | Country Acc (%) | Failure Rate (%) |
|---|---|---|---|---|---|---|---|
| Claude-3.5-S | 56.0 | 17.3 | 12.8 | 3.4 | 31.7 | 61.1 | 0.3 |
| Claude-3.7-S | 60.0 | 19.0 | 12.5 | 4.0 | 32.0 | 62.3 | 0.4 |
| Claude-Opus | 54.0 | 15.2 | 11.3 | 1.0 | 29.1 | 58.5 | 0.3 |
| Claude-S4 | 48.0 | 14.8 | 9.8 | 0.8 | 27.2 | 57.5 | 0.0 |
| GPT-4-T | 60.0 | 21.9 | 16.1 | 4.0 | 35.1 | 65.8 | 0.1 |
| GPT-4.1 | 61.0 | 23.2 | 19.1 | 11.6 | 36.6 | 76.5 | 0.0 |
| GPT-4.1-M | 52.0 | 16.5 | 12.1 | 4.6 | 30.3 | 67.5 | 0.0 |
| GPT-4.1-N | 30.0 | 5.9 | 4.7 | 0.2 | 18.0 | 47.3 | 0.7 |
| GPT-4o | 57.0 | 20.7 | 16.3 | 6.2 | 38.0 | 75.8 | 11.0 |
| GPT-4o-M | 46.0 | 14.8 | 12.3 | 1.2 | 32.9 | 70.8 | 21.7 |
| GPT-o3 | 37.0 | 16.9 | 11.8 | 6.0 | 41.3 | 77.5 | 33.1 |
| GPT-o4-M | 24.0 | 10.5 | 7.9 | 1.2 | 37.8 | 72.0 | 61.4 |
| Gemini-1.5-F | 36.0 | 12.7 | 10.9 | 9.4 | 31.0 | 72.4 | 0.0 |
| Gemini-1.5-8B | 22.0 | 6.8 | 6.6 | 5.0 | 26.3 | 70.9 | 1.4 |
| Gemini-1.5-P | 50.0 | 18.6 | 14.8 | 9.6 | 33.0 | 74.1 | 0.0 |
| Gemini-2.0-F | 42.0 | 16.5 | 10.0 | 4.4 | 32.4 | 70.3 | 7.5 |
| Gemini-2.0-FL | 46.0 | 16.0 | 12.5 | 7.8 | 32.8 | 73.9 | 0.2 |
| Gemma-3-12B | 0.0 | 0.0 | 0.2 | 0.0 | 2.3 | 29.5 | 0.0 |
| Gemma-3-27B-it | 0.0 | 0.0 | 0.2 | 0.0 | 2.3 | 29.5 | 0.0 |
| LLaMA-3.2-11B | 8.0 | 4.2 | 1.8 | 0.0 | 19.1 | 45.3 | 22.6 |
| LLaMA-3.2-90B | 0.0 | 14.8 | 8.7 | 1.6 | 12.8 | 38.1 | 6.8 |
| Qwen-2.5-72B | 36.0 | 10.1 | 6.9 | 2.0 | 29.0 | 65.7 | 0.0 |
| Qwen-2.5-32B | 31.0 | 10.5 | 7.1 | 2.0 | 25.0 | 56.6 | 0.0 |
| Qwen-2.5-7B | 9.0 | 4.2 | 4.3 | 0.8 | 27.3 | 59.0 | 4.3 |
| Idefics-80B | 5.0 | 2.1 | 1.4 | 0.0 | 20.2 | 70.1 | 75.5 |

Table 1: Vision-Language Model Performance Across Geolocation Datasets.

## Experimental Setup

This section details the experimental framework used to evaluate the geolocation capabilities of generative Vision-Language Models (VLMs). We describe the models under study, the datasets used for evaluation, the prompting and data extraction pipeline, and the metrics applied for quantitative and categorical performance analysis.

### Vision-Language Models

We benchmark **25 generative Vision-Language Models (VLMs)**, spanning seven distinct architecture families and parameter scales ranging from 4 billion to over 90 billion parameters (with some closed-source models having undisclosed parameter counts). To facilitate principled and reproducible analysis, we categorize these models into three transparency-based licensing regimes, reflecting their availability and auditability:

- **Closed-source:** Access is restricted to proprietary APIs, with undisclosed model weights, training datasets, and data-filtering heuristics.
- **Open-weight:** Model checkpoints and inference code are publicly available, but training data, preprocessing pipelines, and optimization details remain undisclosed.
- **Open-source:** Full release of model weights, training code, inference code, and comprehensive documentation of the data processing pipeline, enabling complete reproducibility.

**Closed-source models.** We include OpenAI's **GPT-o3, GPT-o4-mini, GPT-4-Turbo, GPT-4o, and GPT-4.1**; Anthropic's **Claude-3.5, Claude-3.7, Claude-4.0, and Claude-Opus**; and Google's **Gemini-1.5 and Gemini-2.0**. While detailed architectures of these systems are proprietary, literature and official documentation indicate they typically employ large-scale transformer architectures, often leveraging mixture-of-experts (MoE) techniques and proprietary visual embedding modules. Prior evaluations (Achiam et al. 2023b; Hochmair, Juhász, and Kemp 2024) confirm state-of-the-art performance in multimodal reasoning tasks, particularly in vision-language reasoning and spatial inference. The inclusion of these models is motivated both by their empirical performance and their potential implications in realistic high-stakes geospatial misuse scenarios.

**Open-weight models.** We evaluate Meta's **LLaMA-3 Vision** models (11B, 90B) (Dubey et al. 2024), Alibaba's **Qwen-2.5-VL** series (7B, 32B, 72B) (Bai et al. 2025), and Google's **Gemma-3** models (4B, 12B) (Team et al. 2025). These models publicly release checkpoints and inference code but obscure critical pre-training details, including visual dataset composition, data augmentation, filtering heuristics, and optimization procedures. Architecturally, these models commonly use unified transformer backbones with learned multimodal embeddings or fusion modules, integrating vision transformer encoders adapted to visual information before language modeling. These open-weight models represent the frontier of publicly accessible checkpoints with strong cross-domain generalization.

**Open-source models.** We benchmark Hugging Face's **IDEFICS** model (80B), which uniquely provides not only checkpoints and inference code but also full training scripts, data preprocessing pipelines, and filtering heuristics under a permissive open-source license (Laurençon et al. 2024). These are the only models in our benchmark allowing full replication from raw input through to final training, providing an important transparency-driven baseline for evaluating reproducibility and auditability.

**Selection rationale.** Our model selection intentionally spans transparency regimes, architectural diversity, and parameter scale to systematically examine how each factor influences model performance and behavior in geolocation-related tasks. We explicitly include models known to have specialized multimodal capabilities (e.g., Claude-Opus and GPT-4o), which reportedly leverage advanced spatial attention mechanisms and long-context reasoning. By including multiple models in each transparency category, we aim to disentangle effects stemming from transparency, modality integration strategies, and model capacity, allowing comprehensive evaluation of risks associated with advanced proprietary systems and publicly available alternatives.

All models are evaluated through a unified inference pipeline to ensure comparability and eliminate confounding effects from interface-specific preprocessing or postprocessing strategies.

## Datasets

We evaluate on four benchmark datasets, each comprising geo-tagged images with precise latitude and longitude annotations. All explicit metadata is removed to ensure predictions are based solely on visual content.

- **IM2GPS** and **IM2GPS3k:** Global Flickr-sourced datasets with diverse natural, rural, and urban outdoor imagery. im2gps includes 237 images, while im2gps3k scales this to 3,000.
- **GPTGeoChat:** A curated dataset of 1,000 social-media-like images, approximately evenly split between U.S. and international locations. Roughly 85% of the images include embedded text, reflecting realistic multimodal content.
- **OSV5:** A street-view dataset spanning 225 countries and 70,000+ cities. For tractability, we randomly sample 500 images, ensuring broad geographic representation, including non-Western and rural areas.

## Prompting and Output Extraction

Each image is presented to the VLM with a standardized prompt designed to elicit structured predictions across multiple geographic levels: country, city, neighborhood, and exact location. The prompt requires complete, reasoned responses in a valid JSON format.

**Standardized Prompt:** *Please provide your speculative guess for the location of the image at the country, city, neighborhood, and exact location levels... [prompt omitted for brevity].*

**Expected JSON Output Format**

```
{
  "rationale": "Country: I chose United
     States because...",
  "country": "United States",
  "city": "New York City",
  "neighborhood": "Manhattan",
  "exact_location_name": "Empire State
     Building",
  "latitude": "40.748817",
  "longitude": "-73.985428"
}
```

**Postprocessing:** We extract structured predictions using regex-based parsing to accommodate minor generation errors. For latitude and longitude, we match coordinates using: `r"-?\d{1,3}.\d+"`. This ensures robust recovery even when the model deviates slightly from strict JSON formatting.

## Evaluation Metrics

Model outputs are evaluated using both continuous geospatial metrics and discrete classification scores.

**Haversine Distance.** The core metric used in our quantitative evaluation is the haversine distance between predicted $(\hat{\phi}, \hat{\lambda})$ and ground-truth $(\phi, \lambda)$ coordinates:

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi)\cos(\hat{\phi})\sin^2\left(\frac{\Delta\lambda}{2}\right)}\right), \tag{1}$$

where $r = 6{,}371$ km is Earth's radius. Here, $\phi$ and $\lambda$ denote latitude and longitude, respectively, given in radians. $\Delta\phi = \hat{\phi} - \phi$ and $\Delta\lambda = \hat{\lambda} - \lambda$ represent the differences between the predicted and ground-truth latitude and longitude.

**Recall@Nkm.** We compute Recall@Nkm, the proportion of predictions within $N$ km haversine distance of ground truth:

$$\text{Recall@}N = \frac{1}{M}\sum_{i=1}^{M} \mathbb{I}[d_i \leq N], \tag{2}$$

reporting Recall@1, 25, 200, and 750 km to cover street-, city-, regional-, and country-level granularity.

**Administrative Accuracy.** We report classification accuracy at the country and city levels:

$$\text{Accuracy}_{\text{country/city}} = \frac{1}{M}\sum_{i=1}^{M} \mathbb{I}[\hat{c}_i = c_i] \tag{3}$$

All labels are normalized using services like GeoNames or OpenStreetMap to resolve spelling variants and naming inconsistencies.

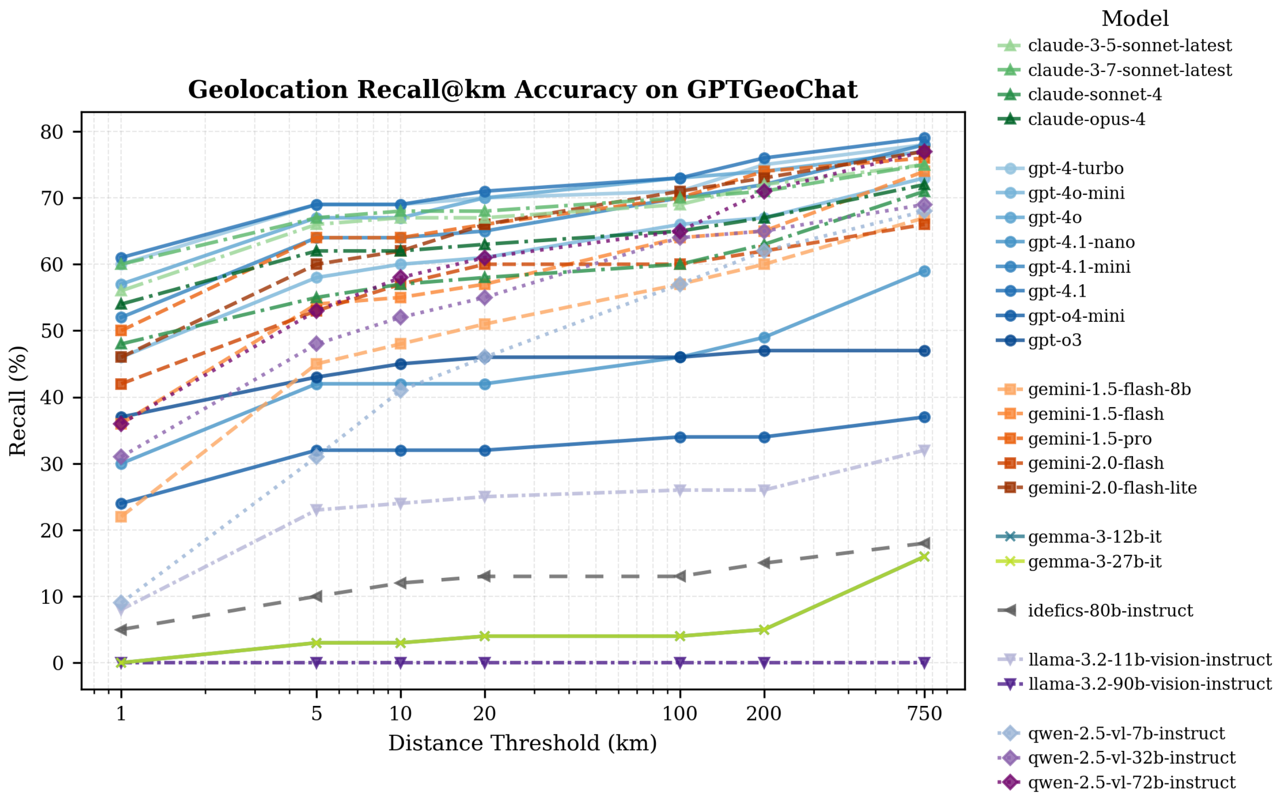


Figure 2: Recall@km performance of 25 open-weight, and open- and closed-source generative VLMs evaluated on the GPTGeoChat dataset.

This mixed evaluation approach quantifies both spatial proximity and semantic localization fidelity, offering a comprehensive assessment of model geolocation capabilities.

## Results and Analysis

This section analyzes the geolocation capabilities of the evaluated VLMs across three aspects: *Quantitative Localization Accuracy* which focuses on distance-based Recall@km metrics; *Administrative Region Accuracy*, where we evaluate categorical predictions at city and country levels; and *Failure Analysis*, where we evaluate the ability of the models to provide a geolocation estimation.

### Quantitative Localization Accuracy

Table 1 summarizes quantitative results across the four benchmark datasets, with figure 2 showing the Recall@km accuracy across distances on the GPTGeoChat dataset. Overall, GPT-4.1 consistently emerges as the strongest performer, achieving the highest recall across all evaluated distance thresholds (1km, 25km, 200km, and 750km). Specifically, GPT-4.1 attains a remarkable Recall@1km of 61% on GPTGeoChat, outperforming other closed-source models such as Claude-3.7-Sonnet (60%) and Gemini-1.5-Pro (50%).

In comparison, open-weight models exhibit a notable gap in performance. The best-performing open-weight model, Qwen-2.5-72B, achieves a Recall@1km of 36% on GPTGeoChat, indicating less precise geospatial understanding. Meanwhile, fully open-source models (IDEFICS) show limited geolocation capabilities, with a maximum Recall@1km of only 5%. This modest but measurable performance indicates that existing data mixtures used in its training, such as Wikipedia, OBELICS, and LAION, provide some foundational support for geolocation tasks albeit with relatively low precision.

Performance varies considerably depending on dataset characteristics. Closed-source models perform exceptionally well on GPTGeoChat, which closely mimics social media imagery containing distinctive visual and textual cues. In contrast, performance drops substantially for datasets such as OSV5, characterized by broader geographic diversity and fewer recognizable visual landmarks. Here, GPT-4.1 again leads but with a much lower Recall@100km of 11.6%, reflecting difficulty in generalizing across geographically diverse imagery. While we describe these as model capabilities, such patterns likely arise from the composition and biases of the training corpus rather than inherent architectural properties. Notably, models from the Gemma family consistently failed to produce direct geographic predictions, significantly limiting their applicability for geolocation tasks.

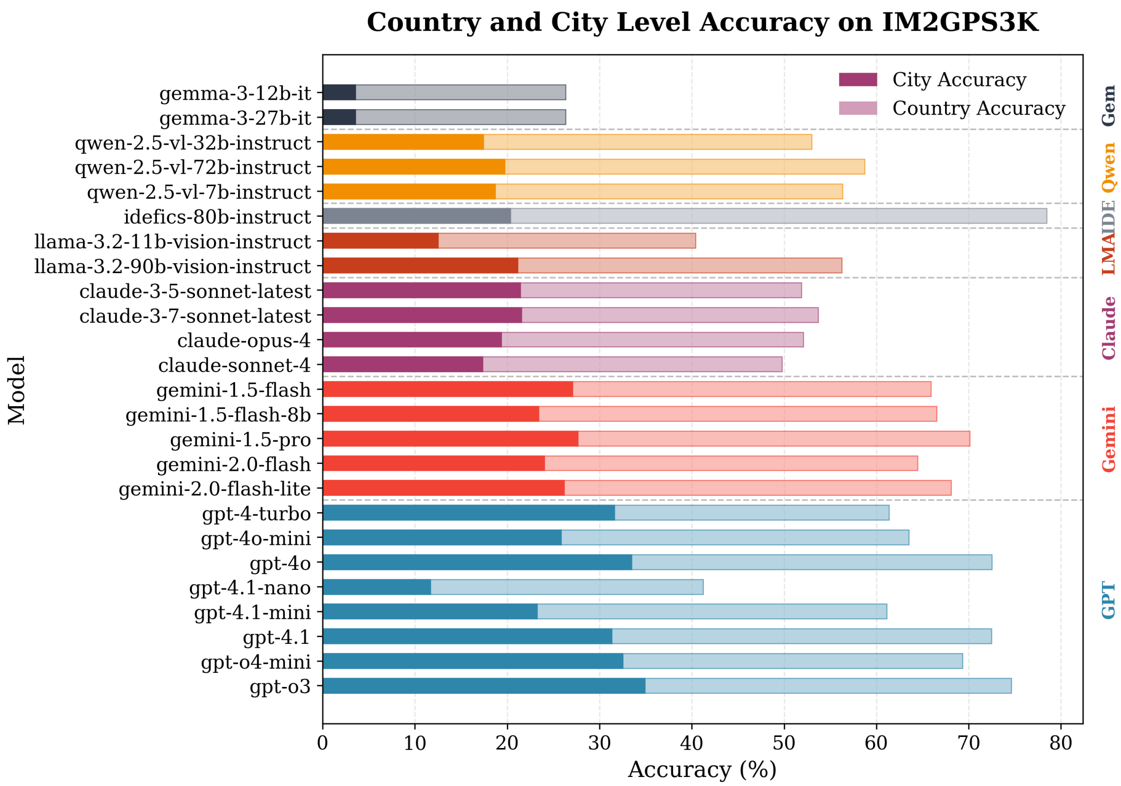


Figure 3: Categorical accuracy (city- and country-level) of 25 generative VLMs evaluated on the IM2GPS3k dataset.

## Administrative Region Accuracy

Figure 3 shows the categorical accuracy of model predictions at administrative levels, specifically country and city accuracy, for im2gps3k. Unlike distance-based metrics, administrative accuracy evaluates the semantic correctness of geographic labels independently from precise numeric coordinates.

Table 1 summarizes these results averaged across all datasets. GPT-o3 attains the highest overall performance, with mean accuracies of 41.3% at the city level and 77.5% at the country level. GPT-4o follows closely (38.0% city, 75.8% country), indicating that both models maintain strong semantic understanding of geographic regions. Among open-weight models, performance is generally lower but remains consistent with model scale: Qwen-2.5-72B achieves 29.0% city-level and 65.7% country-level accuracy, outperforming smaller counterparts. Among open-source models, IDEFICS surprisingly attains 20.2% city-level and 70.1% country-level accuracy, suggesting that it retains nontrivial geographic knowledge.

## Failure Analysis

We observe varied failure modes across evaluated models in figure 4. Notably, certain models demonstrate high refusal rates or frequently omit numeric predictions entirely. IDEFICS shows the highest overall failure rate (75.5%), primarily attributable to refusal to generate structured numeric coordinates. Other models, such as GPT-o4-Mini, exhibit intermittent failures (61.4%), largely caused by refusals to answer, such as regularly responding with "*I'm sorry, but I can't help with that.*"

These failures underscore critical reliability challenges for real-world applications, particularly where structured geolocation output is essential. However, analyzing these failures also helps understand the safety profile and potential societal risks models like GPT-4.1 may pose.

# Discussion and Implications

Our evaluation highlights a clear asymmetry in the geolocalization capabilities of VLMs. While we report these as capabilities and limitations of the models, many of these patterns likely originate from the data they were trained on; the models perform poorly on structured imagery such as street views but show striking accuracy on social media–like content. This disparity suggests an underlying bias: current VLMs are more effective at identifying people-centric environments than broader geographic locations. This misalignment raises important concerns regarding applications like disaster response, urban planning, and environmental monitoring, which depend on accurate localization across a wide range of scenes that may lack distinctive visual cues. The models' limited performance on these inputs calls into question their suitability for these socially beneficial use cases. Instead, their strengths appear disproportionately concentrated on identifying individuals and their surroundings, whether or not that is the intended goal.

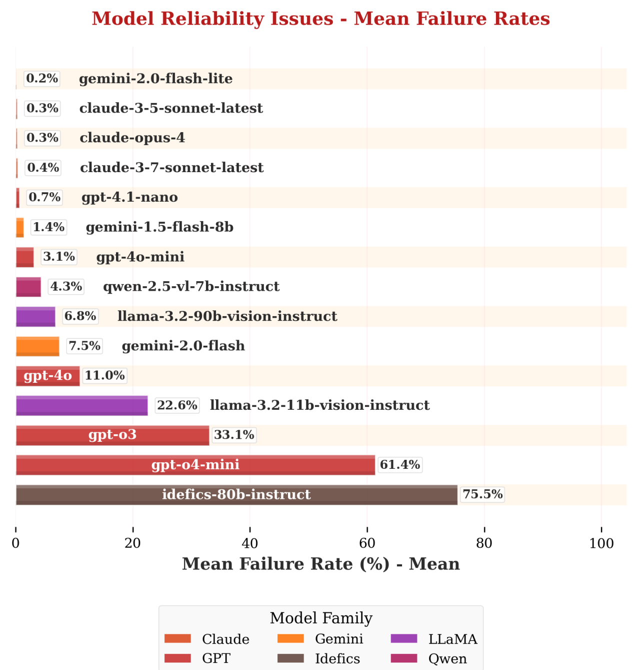


Figure 4: Failure rates including both refusal to respond and failure to produce numeric predictions.

These capabilities have direct privacy implications. Visual content shared online, even when stripped of explicit location markers, can be reverse-engineered to reveal precise locations. Users typically lack both awareness and control over how such inferences are made. As models improve and become widely accessible, the potential for privacy breaches becomes more pervasive and difficult to guard against. There are predictable threat models, including stalking, surveillance and targeting, which would be heightened by powerful expert users in autocratic regimes. But we can also imagine economically profitable use cases in which companies monetise image geolocation as a new source of personal data for advertising and other forms of what (Hongladarom 2020) calls 'surveillance capitalism'. Geolocation could become part of the future political economy of social media.

Ultimately, these findings underscore the need for ethical safeguards, transparent model behavior, and better user protections. As VLMs continue to evolve, understanding and addressing their sociotechnical risks will be critical to ensure that progress in capability does not come at the expense of privacy and safety.

## Conclusion

We conducted a comprehensive evaluation of the geolocalization capabilities of state-of-the-art generative VLMs, comparing closed-source, open-weight, and fully open-source systems across diverse benchmark datasets. Using both distance-based and administrative-level metrics, we found that nearly all models achieve non-negligible accuracy, with closed-source models leading overall performance. Performance varied considerably with dataset characteristics, with models excelling on social media–like images and struggling on less visually distinctive scenes. These results demonstrate that current VLMs already possess sufficient geo-localization ability to pose considerable privacy risks, particularly when applied to publicly shared visual content.

## Acknowledgments

This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) and the UK AI Security Institute (AISI) through grant (PRIV-LOC: Assessing and Mitigating Privacy Risks of Vision-Language Models in Image-based Geolocation Systems). It is also supported in part by the Responsible AI UK programme, EPSRC grant number EP/Y009800/1.

## References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023a. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023b. Gpt-4 technical report.

Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; and Sivic, J. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5297–5307.

Ashbrook, D.; and Starner, T. 2003. Starner, T.: Using GPS to learn significant locations and predict movement across multiple users. Personal and Ubiquitous Computing 7(5), 275-286. *Personal and Ubiquitous Computing*, 7: 275–286.

Astruc, G.; Dufour, N.; Siglidis, I.; Aronssohn, C.; Bouia, N.; Fu, S.; Loiseau, R.; Nguyen, V. N.; Raude, C.; Vincent, E.; Xu, L.; Zhou, H.; and Landrieu, L. 2024. OpenStreetView-5M: The Many Roads to Global Visual Geolocation. *CVPR*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Carlini, N.; Ippolito, D.; Jagielski, M.; Lee, K.; Tramèr, F.; and Zhang, C. 2022. Quantifying Memorization Across Neural Language Models. *arXiv:2202.07646*.

Carlini, N.; Tramèr, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T. B.; Song, D. X.; Erlingsson, Ú.; Oprea, A.; and Raffel, C. 2020. Extracting Training Data from Large Language Models. In *USENIX Security Symposium*.

Chen, N. C.; Xie, W.; Welsch, R. E.; Larson, K.; and Xie, J. 2017. Comprehensive Predictions of Tourists' Next Visit Location Based on Call Detail Records Using Machine Learning and Deep Learning Methods. In *2017 IEEE International Congress on Big Data (BigData Congress)*, 1–6.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Gerber, N.; Gerber, P.; and Volkamer, M. 2018. Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior. *Computers & Security*, 77: 226–261.

Haas, L.; Alberti, S.; and Skreta, M. 2023. Learning Generalized Zero-Shot Learners for Open-Domain Image Geolocalization. arXiv:2302.00275.

Haas, L.; Skreta, M.; Alberti, S.; and Finn, C. 2024. PIGEON: Predicting Image Geolocations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12893–12902.

Hargittai, E.; and Marwick, A. 2016. "What Can I Really Do?" Explaining the Privacy Paradox with Online Apathy. *International Journal of Communication*, 10(0): 21.

Hays, J.; and Efros, A. A. 2008. Im2gps: estimating geographic information from a single image. In *2008 ieee conference on computer vision and pattern recognition*, 1–8. IEEE.

Hochmair, H. H.; Juhász, L.; and Kemp, T. 2024. Correctness comparison of ChatGPT-4, Gemini, Claude-3, and Copilot for spatial tasks. *Transactions in GIS*, 28(7): 2219–2231.

Hongladarom, S. 2020. Shoshana Zuboff, The age of surveillance capitalism: the fight for a human future at the new frontier of power: New York: Public Affairs, 2019, 704 pp. ISBN 978-1-61039-569-4 (hardcover) 978-1-61039-270-0 (ebook). *AI Soc.*, 38(6): 2359–2361.

Kokolakis, S. 2017. Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers & Security*, 64: 122–134.

Kordopatis-Zilos, G.; Galopoulos, P.; Papadopoulos, S.; and Kompatsiaris, I. 2021. Leveraging efficientnet and contrastive learning for accurate global-scale location estimation. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, 155–163.

Krumm, J. 2022. Sensitivity Analysis of Personal Location Disclosure. In *2022 23rd IEEE International Conference on Mobile Data Management (MDM)*, 73–82.

Laurençon, H.; Tronchon, L.; Cord, M.; and Sanh, V. 2024. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37: 87874–87907.

Lazar, S.; and Nelson, A. 2023. AI safety on whose terms? *Science*, 381(6654): 138–138.

Luo, G.; Biamby, G.; Darrell, T.; Fried, D.; and Rohrbach, A. 2022. $G^3$: Geolocation via Guidebook Grounding. arXiv:2211.15521.

Marmasse, N.; and Schmandt, C. 2002. A User-Centered Location Model. *Personal Ubiquitous Comput.*, 6(5–6): 318–321.

Mendes, E.; Chen, Y.; Hays, J.; Das, S.; Xu, W.; and Ritter, A. 2024. Granular Privacy Control for Geolocation with Vision Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 17240–17292. Miami, Florida, USA: Association for Computational Linguistics.

Muller-Budack, E.; Pustu-Iren, K.; and Ewerth, R. 2018. Geolocation estimation of photos using a hierarchical model and scene classification. In *Proceedings of the European conference on computer vision (ECCV)*, 563–579.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

Solove, D. J. 2025. On Privacy and Technology (excerpt). https://ssrn.com/abstract=5159448. Daniel J. Solove, *On Privacy and Technology* (Oxford University Press, 2025).

Team, G.; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; et al. 2025. Gemma 3 technical report.

Vivanco, V.; Nayak, G. K.; and Shah, M. 2023. GeoCLIP: Clip-Inspired Alignment between Locations and Images for Effective Worldwide Geo-localization. In *Advances in Neural Information Processing Systems*.

Vo, N.; Jacobs, N.; and Hays, J. 2017. Revisiting IM2GPS in the Deep Learning Era. arXiv:1705.04838.

Waheed, S.; Ferrarini, B.; Milford, M.; Ramchurn, S. D.; and Ehsan, S. 2025. Image-based Geo-localization for Robotics: Are Black-box Vision-Language Models there yet? arXiv:2501.16947.

Weyand, T.; Araujo, A.; Cao, B.; and Sim, J. 2020. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2575–2584.

Weyand, T.; Kostrikov, I.; and Philbin, J. 2016. Planet-photo geolocation with convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, 37–55. Springer.

Wu, M.; and Huang, Q. 2022. IM2City: image geo-localization via multi-modal learning. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, GeoAI '22, 50–61. New York, NY, USA: Association for Computing Machinery. ISBN 9781450395328.

Zhou, Z.; Zhang, J.; Guan, Z.; Hu, M.; Lao, N.; Mu, L.; Li, S.; and Mai, G. 2024. Img2Loc: Revisiting Image Geolocalization using Multi-modality Foundation Models and Image-based Retrieval-Augmented Generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, volume 35 of *SIGIR 2024*, 2749–2754. ACM.