# Interim Report on an Image Geolocation Research Project

Hoffmann Szabolcs

May 17, 2026

## 1 Introduction

The aim of this project is to investigate image geolocation, that is, the task of predicting the geographical coordinates at which an image was taken. The long-term objective is to reproduce, at least partially, the methodology of PIGEON, a recent image geolocation system designed for planet-scale location prediction. PIGEON combines modern vision-language models, semantic geocells, haversine-based evaluation and retrieval-based refinement.

At the current stage, the full recreation of PIGEON has not yet been completed. The main reason is computational: a meaningful reproduction requires GPU access for training or fine-tuning image models. Since access to a suitable GPU system has only recently become available, the actual PIGEON-style model training will be continued in the next semester. Therefore, the present phase focused on understanding the paper, exploring possible datasets, building a data and evaluation pipeline, and running a preliminary benchmark using ChatGPT as a zero-shot image geolocation model.

A further motivation for testing ChatGPT is that it has capabilities that PIGEON does not explicitly have. In particular, ChatGPT can often read visible text in images, such as signs, shop names, public notices, road signs, inscriptions or banners. This OCR-like ability can be highly relevant for geolocation, since textual clues often reveal the language, country, city or even exact place. PIGEON, in contrast, is primarily a specialised visual geolocation model and does not explicitly use OCR as a separate component. Thus, the ChatGPT experiment is not only a temporary substitute before GPU-based reproduction, but also an exploratory step toward a possible hybrid system in which a PIGEON-like model and an OCR-capable multimodal model help each other.

## 2 Dataset Search and Data Preparation

A central difficulty in this project is the choice of dataset. The ideal dataset would contain many images with accurate latitude and longitude values. For a PIGEON-style reproduction, street-level or panoramic data would be best. However, the dataset used by the original PIGEON authors is not directly available, and other promising sources such as Mapillary require API access tokens.

Several alternatives were considered. Mapillary is highly relevant because it contains street-level geotagged images, but it requires access through its API. Google Landmarks v2 is useful for landmark recognition, but it is not primarily an exact-coordinate benchmark. Wikimedia Commons, on the other hand, contains freely accessible geotagged images and can be queried without a private API token. For this reason, Wikimedia Commons was chosen for the preliminary benchmark.

A Python script was written to collect geotagged Wikimedia Commons images around selected global seed points. The script downloaded the images, saved their true latitude and longitude coordinates, and grouped the images into batches of ten. The true coordinates were stored separately from the images, so that they would not be visible during ChatGPT prediction.

In total, 169 images were collected, approximately 30 per region. However, not all of them were suitable for the intended evaluation. Some images showed people, indoor spaces, museum interiors, objects, or scenes with very weak geographical information. Since the aim was to evaluate outdoor visual geolocation, manual filtering was introduced. Images judged unsuitable were removed manually, and the evaluation was later performed only on the remaining usable images.

## 3 ChatGPT-Based Benchmark

Before training a PIGEON-like model, a preliminary benchmark was built using ChatGPT as a zero-shot geolocation predictor. The workflow was:

$$\text{image} \rightarrow \text{ChatGPT coordinate prediction} \rightarrow \text{haversine error.}$$

The images were uploaded manually to ChatGPT Pro in batches of ten. For each batch, ChatGPT was asked to return a CSV table with the columns

`image_id`, `predicted_lat`, `predicted_lon`, `confidence_0_1`, `country_guess`, `short_reason`.

The prompt explicitly asked for one numerical coordinate estimate per image, even if the model was uncertain. The predictions were then cleaned, merged with the hidden ground-truth coordinates, and evaluated automatically.

The first successful evaluation contained 54 usable predictions. The main results were as follows:

| Metric | Value |
|---|---|
| Number of evaluated images | 54 |
| Mean error | 304.17 km |
| Median error | 3.58 km |
| Minimum error | 0.003 km |
| Maximum error | 5235.18 km |
| % within 1 km | 29.63% |
| % within 25 km | 87.04% |
| % within 200 km | 88.89% |
| % within 750 km | 92.59% |
| % within 2500 km | 94.44% |

These results should be interpreted carefully. The median error is very low, which suggests that many images were easy to identify, probably because they contained famous landmarks, recognisable urban scenes or readable text. At the same time, the mean error is much higher because a few predictions failed by several thousand kilometres. This is typical for geolocation tasks: a model may often be close, but occasional continent-level mistakes strongly affect the average.

The results also support the idea that OCR-like reasoning can be useful. In several cases, ChatGPT appeared to use visible text, place names, inscriptions or culturally specific signs to improve its guess. This suggests that ChatGPT and PIGEON may have complementary strengths. PIGEON-like models may be better at learning large-scale visual distributions, while ChatGPT may be particularly useful when textual or semantic clues are visible.

# 4 Evaluation Method

The evaluation uses haversine distance, which measures the great-circle distance between two latitude–longitude points on the surface of the Earth. If the true location is

$$p = (\varphi_1, \lambda_1)$$

and the predicted location is

$$\hat{p} = (\varphi_2, \lambda_2),$$

then the haversine distance is

$$d(p, \hat{p}) = 2R \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right),$$

where $R$ is the Earth's radius, approximately 6371 km.

In addition to the raw distance error for each image, the project computes the mean error, median error, minimum and maximum error, and the percentage of predictions within 1 km, 25 km, 200 km, 750 km and 2500 km. These thresholds are useful because they correspond roughly to street-level, city-level, regional, country-level and continent-level accuracy.

Several scripts were implemented for this process. One script downloads the images and ground-truth coordinates, another cleans the ChatGPT prediction files, a third filters the ground-truth table after manual image deletion, and a fourth computes all error statistics. The results were also exported to an Excel file with separate sheets for summary statistics, individual predictions and the largest errors.

# 5 Limitations and Future Work

The current benchmark is only a preliminary experiment. The sample size is small, and the Wikimedia Commons images are not representative of arbitrary street-level locations. Many of them are landmarks or culturally significant places, which may make them easier for ChatGPT than random locations. Manual filtering also introduces subjectivity, although it is justified by the goal of focusing on outdoor geolocation.

Another limitation is that ChatGPT and PIGEON do not solve exactly the same task in the same way. ChatGPT may exploit OCR-like recognition of visible text, while PIGEON is primarily a specialised visual model. This makes direct comparison difficult, but it is also one of the most interesting aspects of the project. A future hybrid system could use a PIGEON-like model to provide an initial distribution over likely regions, and then use ChatGPT or another OCR-capable multimodal model to refine the prediction based on text, language and semantic clues.

The next semester will focus on the actual PIGEON reproduction, now that GPU access has become available. The main planned steps are:

- setting up a reproducible GPU environment;
- studying and adapting the available PIGEON codebase;
- building or obtaining a larger image–coordinate dataset;
- implementing a baseline geocell classifier;

- evaluating it with the same haversine-based metrics;
- comparing the trained model with the current ChatGPT benchmark.

I will also try to contact the authors or owners of the original PIGEON project and request access to the dataset on which they tested their model, or at least to a comparable evaluation subset. If this is possible, the comparison between ChatGPT, the reproduced PIGEON-like system and a possible hybrid approach would become much more rigorous.

# 6 Conclusion

The full recreation of PIGEON has not yet been completed, but the project has made significant preparatory progress. The methodology of PIGEON has been studied, possible datasets have been explored, a token-free geotagged image collection pipeline has been implemented, and a first ChatGPT-based image geolocation benchmark has been successfully run.

The preliminary results show that ChatGPT can perform surprisingly well on some geotagged outdoor images, especially when landmarks or textual cues are present. However, the large maximum error and the difference between mean and median error show that zero-shot geolocation remains unstable. This confirms the need for a specialised PIGEON-like model in the next phase.

Overall, the current work should be viewed as an exploratory and preparatory stage. It produced a functioning evaluation framework and suggested a promising future direction: combining specialised visual geolocation models with OCR-capable multimodal models. Such a hybrid approach may eventually be stronger than either component alone.