CRCV REU Week #7

Project: Geo-HiPO

Advisors:

- Rohit Gupta
- Parth Kulkarni
- Dr. Shah

David Rivard

07/11/2025

Project Overview



Motivation:

- Relative spatial reasoning as new fine-tuning strategy for VLMs.
  - By introducing preference pairs.
- VLMs won't have to be retrained on massive geo-tagged dataset.
  - Should generalize to new scenes better.

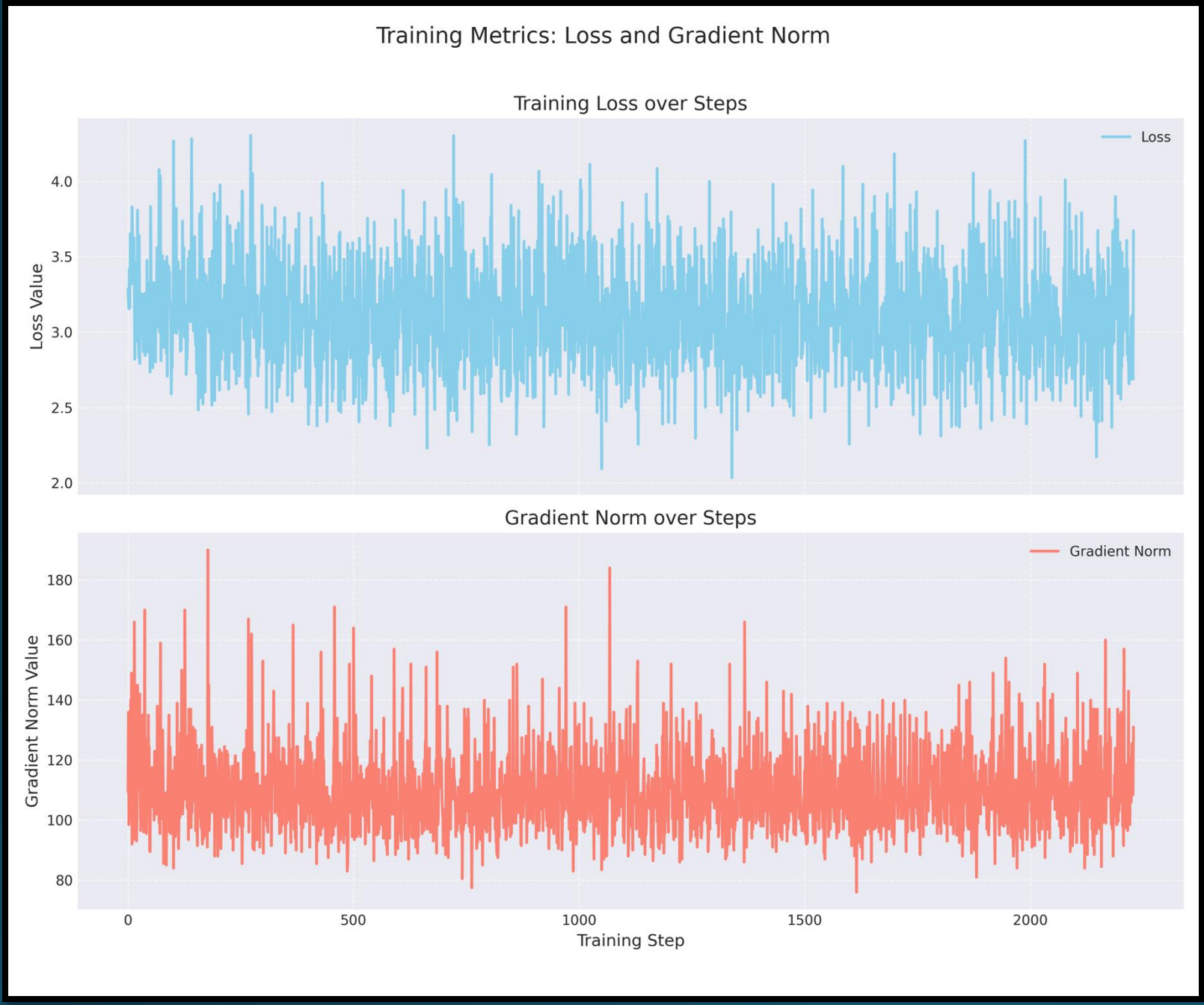
Research Question:

- Can we improve geolocation accuracy by **teaching vision-language models to reason location like humans**?

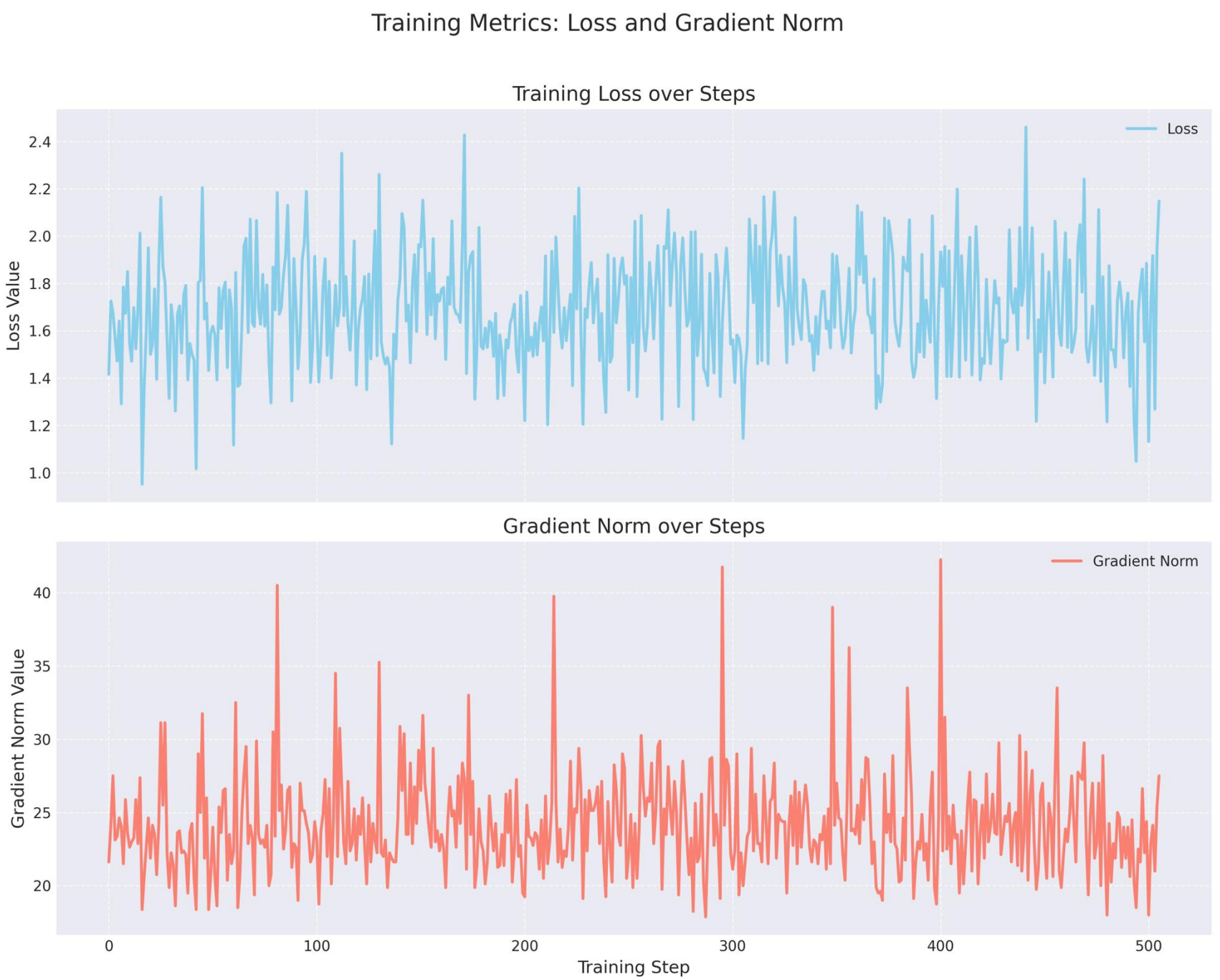
Key Progress

- Lots of Failed Training
- Found a Repo for DPO
  - Starting to integrate it
- Filtered Training Samples
  - Performing SFT on gold samples
- Finalized Training Pair Style
- Running Multiple Trainings

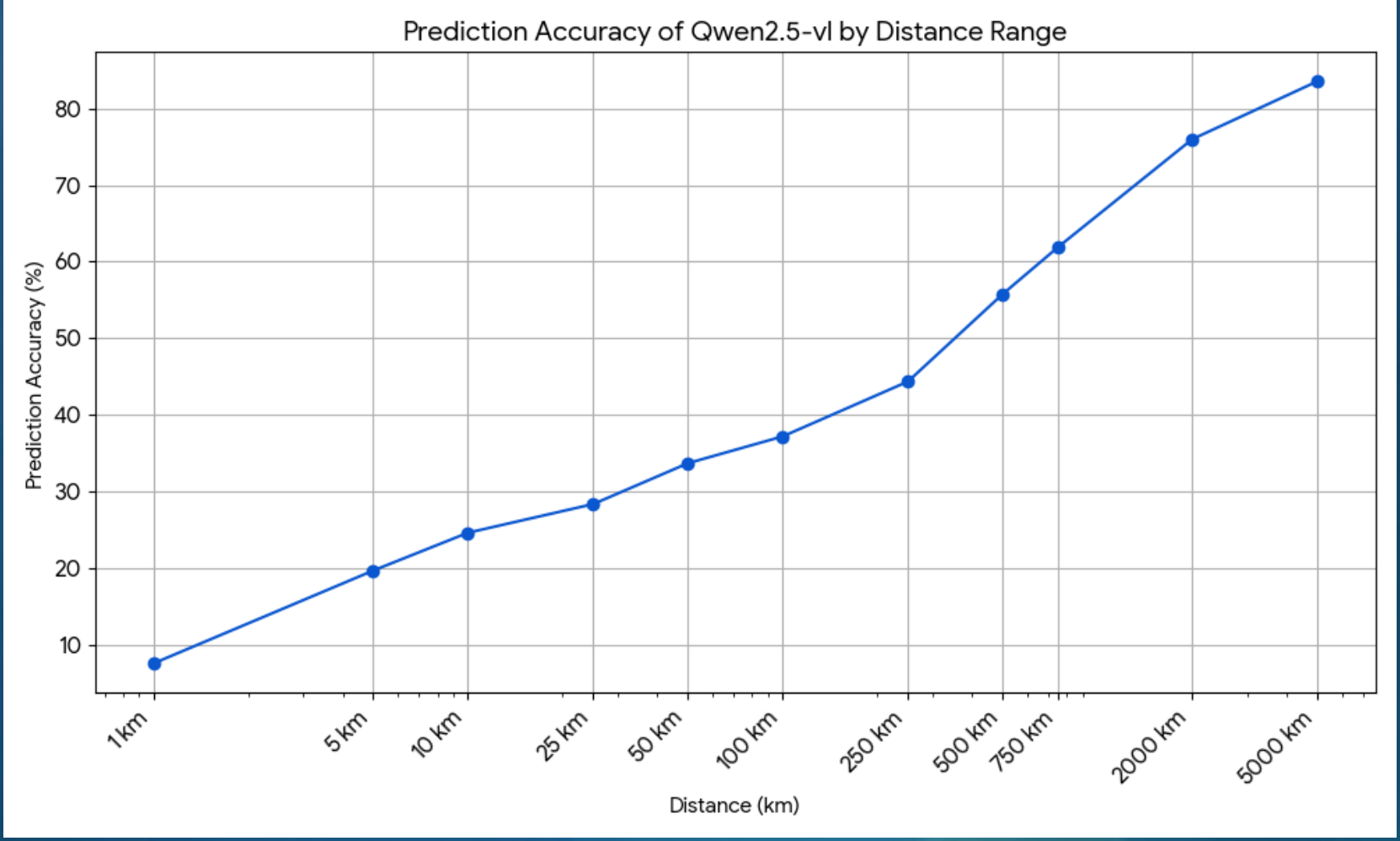
Failed Training



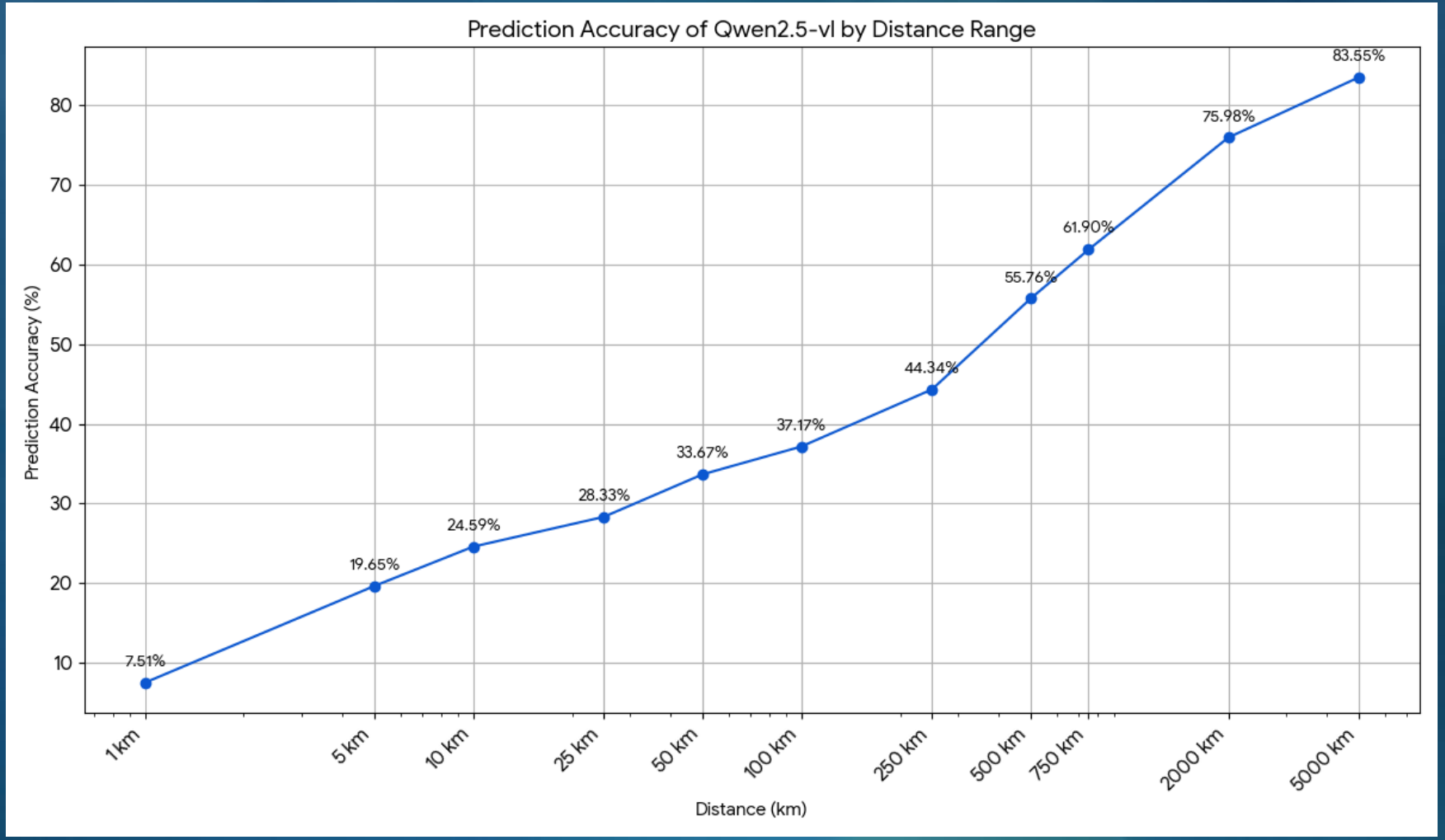
Failed Training



Baseline


Prediction Accuracy of Qwen2.5-vl by Distance Range
Prediction Accuracy (%)
10
20
30
40
50
60
70
80
1 km
5 km
10 km
25 km
50 km
100 km
250 km
500 km
750 km
2000 km
5000 km
Distance (km)

Bad Config
-After Training



Tuning Parameters

--tune_mm_vision True

--tune_mm_mlp True

--tune_mm_llm True

DPO

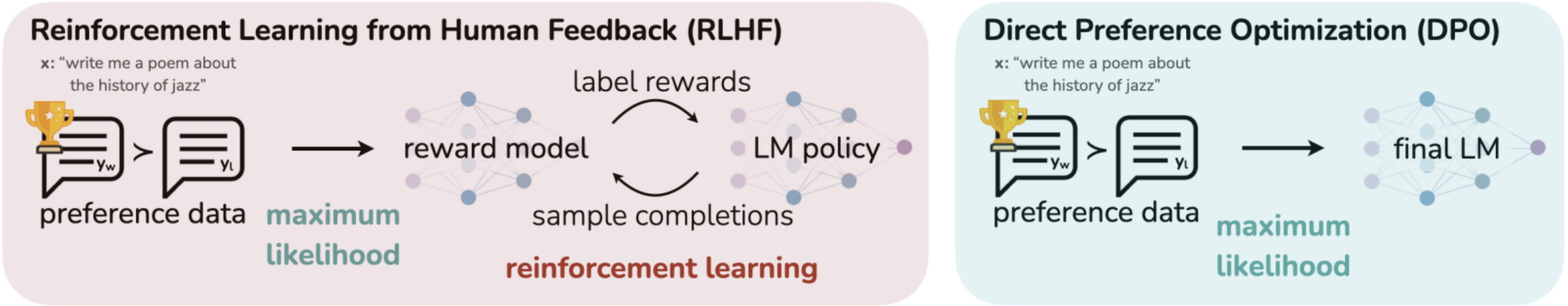
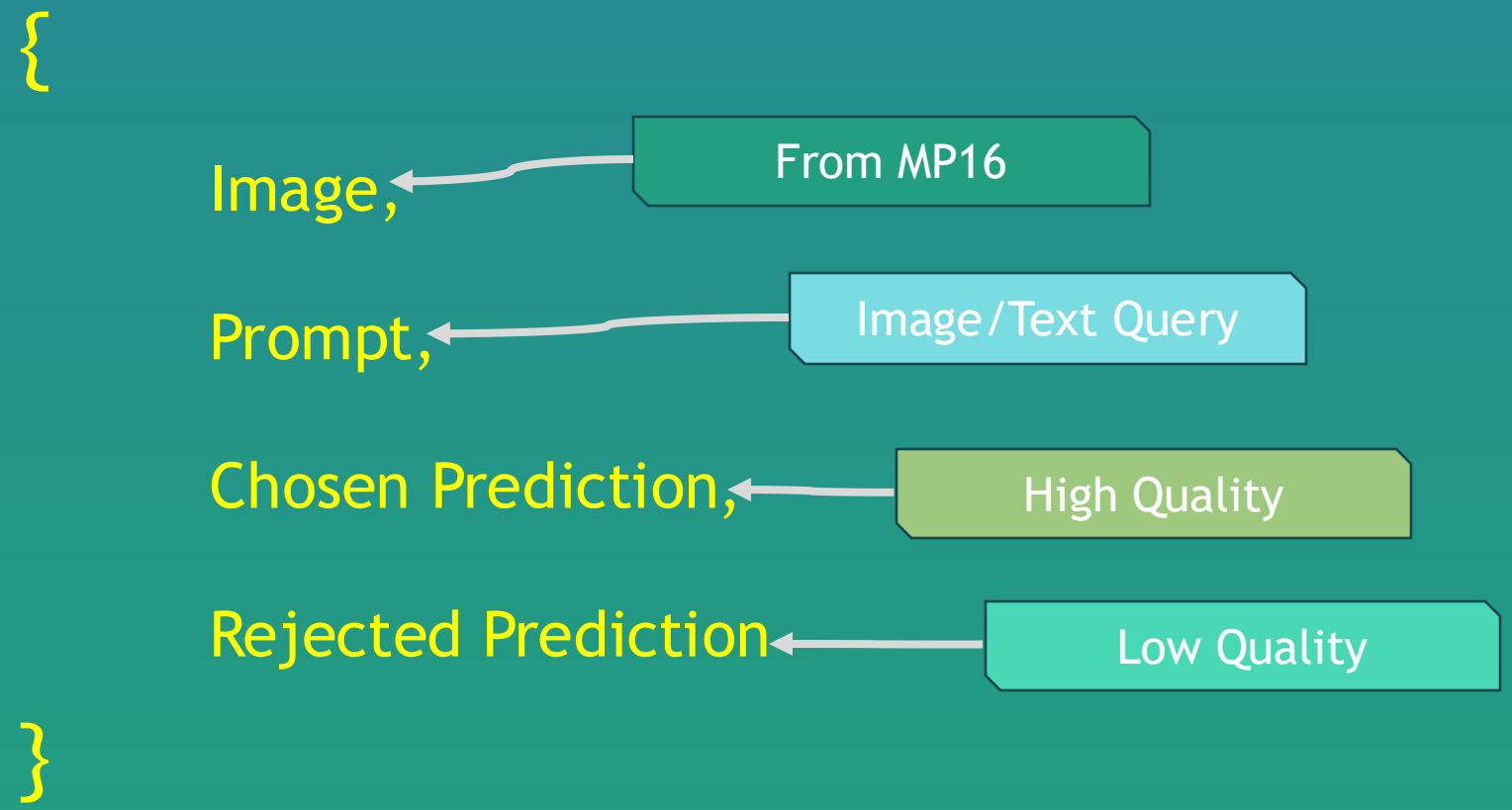


Figure 1: **DPO optimizes for human preferences while avoiding reinforcement learning.** Existing methods for fine-tuning language models with human feedback first fit a reward model to a dataset of prompts and human preferences over pairs of responses, and then use RL to find a policy that maximizes the learned reward. In contrast, DPO directly optimizes for the policy best satisfying the preferences with a simple classification objective, fitting an *implicit* reward model whose corresponding optimal policy can be extracted in closed form.

Preference Pairs



Preference Pairs

Current:


```
{
    "filename": "917538459046089.jpg",
    "closer": "Ashgabat, Turkmenistan",
    "farther": "Astana, Kazakhstan",
    "ground_truth": "37.895893, 58.318624"
  }
```

Reasoning

“”

The image features a Gothic-style church with two prominent towers, each topped with a sharp, conical spire. The architectural style is characterized by pointed arches, ribbed vaults, and ornate details typical of Gothic churches. The building is constructed from light-colored stone, suggesting it might be located in a region where limestone or similar materials are readily available. There are no visible natural elements like trees in the foreground, but greenery can be seen around the base of the church, indicating some landscaping. No signs, vehicles, or infrastructure are visible in the image.\n\nCity, Country\nLoir-et-Cher, France

“”

Plan


Finalize Report & Paper
Evaluation & Refinement
Implement HiPO
~~Filter Training Samples~~
~~Setup vLLM~~
~~Deeper Model Inference~~
~~Baseline Inference~~
~~Benchmark Datasets~~
~~Environment Setup~~
~~Research Papers~~

Thank You

Questions ?