Instruction-Tuning Llama-3-8B Excels in City-Scale Mobility Prediction

Peizhi Tang
Southern Univ. of Sci. and Tech.
Shenzhen, China
12110502@mail.sustech.edu.cn

Chuang Yang*
The Univerisity of Tokyo
Tokyo, Japan
chuangyang@g.ecc.u-tokyo.ac.jp

Tong Xing
The University of Tokyo
Tokyo, Japan
tongxing707@g.ecc.u-tokyo.ac.jp

Xiaohang Xu
The University of Tokyo
Tokyo, Japan
xhxu@g.ecc.u-tokyo.ac.jp

Renhe Jiang
The University of Tokyo
Kashiwa, Japan
jiangrh@csis.u-tokyo.ac.jp

Kaoru Sezaki
The University of Tokyo
Tokyo, Japan
sezaki@iis.u-tokyo.ac.jp

Abstract

Human mobility prediction plays a critical role in applications such as disaster response, urban planning, and epidemic forecasting. Traditional methods often rely on designing crafted, domain-specific models, and typically focus on short-term predictions, which struggle to generalize across diverse urban environments. In this study, we introduce Llama-3-8B-Mob, a large language model fine-tuned with instruction tuning, for long-term citywide mobility prediction—in a Q&A manner. We validate our approach using large-scale human mobility data from four metropolitan areas in Japan, focusing on predicting individual trajectories over the next 15 days. The results demonstrate that Llama-3-8B-Mob excels in modeling long-term human mobility—surpassing the state-of-the-art on multiple prediction metrics. It also displays strong zero-shot generalization capabilities—effectively generalizing to other cities even when fine-tuned only on limited samples from a single city. Source codes are available at https://github.com/TANGHULU6/Llama3-8B-Mob.

CCS Concepts

• **Human-centered computing → Mobile computing**; • **Computing methodologies → Natural language generation**.

Keywords

Human Mobility, Large Language Model, Long-term Forecasting, Transfer Learning

ACM Reference Format:
Peizhi Tang, Chuang Yang, Tong Xing, Xiaohang Xu, Renhe Jiang, and Kaoru Sezaki. 2024. Instruction-Tuning Llama-3-8B Excels in City-Scale Mobility Prediction. In *2nd ACM SIGSPATIAL International Workshop on the Human Mobility Prediction Challenge (HuMob'24), October 29-November 1 2024, Atlanta, GA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3681771.3699908

*Corresponding Author and Project Leader.

HuMob'24, October 29-November 1 2024, Atlanta, GA, USA

ACM ISBN 979-8-4007-1150-3/24/10
https://doi.org/10.1145/3681771.3699908

1 Introduction

Human mobility prediction plays a vital role in various critical scenarios, such as disaster management [1], epidemic forecasting [2], and personalized location recommendations [3]. Numerous advanced machine learning models have been proposed for individual-level human mobility prediction. For instance, Feng et al. proposed a model that integrates Recurrent Neural Networks (RNNs) with attention mechanisms to jointly capture users' long- and short-term preferences for next-location prediction [4]. Similarly, Xu et al. transformed both individual historical movements and collective migration patterns into graphs, applying graph learning to capture complex spatial-temporal trends, leading to significant improvements in prediction accuracy [3]. However, these approaches rely on designing crafted, domain-specific models, which often struggle to generalize effectively across different cities and applications.

In recent years, Large Language Models (LLMs) have rapidly emerged, excelling not only in Natural Language Processing (NLP) tasks such as translation, sentiment analysis, and entity extraction but also redefining various scenarios like data preprocessing [5] and time series prediction [6]. In this context, human mobility modeling is also experiencing a significant paradigm shift. Wang et al. designed an effective prompting strategy that enables LLMs to perform zero-shot next-location prediction in a Q&A format for the first time [7]. The results showed that LLMs demonstrated competitive prediction performance and could offer better interpretability compared to existing domain-specific machine learning models. Furthermore, Li et al. fine-tuned the Llama-2-7B model by framing an individual's historical trajectory as input and the next location as the response, achieving state-of-the-art (SOTA) on three real-world check-in datasets [8]. These studies highlight the enormous potential of LLMs in human mobility modeling. However, several limitations remain: most work just focused on short-term predictions (i.e., next-location prediction), leaving long-term prediction largely unexplored, and the transferability among cities has yet to be examined, making it unclear whether mobility patterns learned in one city can be effectively transferred to another.

To this end, we introduce **Llama-3-8B-Mob**, an instruction-tuned version of Llama-3-8B [9], designed for long-term, multi-city human mobility prediction. We validate our approach on a large-scale human trajectory dataset from four metropolitan areas in Japan, focusing on predicting human mobility over the next 15 days. The results significantly surpass the existing SOTA method, showcasing the impressive performance of Llama-3-8B-Mob in long-term prediction and its strong zero-shot generalization ability—training in a single city effectively generalizes to others. In the Human Mobility Prediction Challenge 2024[1], Llama-3-8B-Mob ranked among the top 10, surpassing over 100 competing prediction models.

2 Problem Definition

Definition 1. (**Trajectory**): A trajectory for a moving object consists of a sequence of spatial-temporal records that depict the object's location at a specific timestamp. Formally, it is defined as:

$$T = (uid, \{(t_i, x_i, y_i) | i = 1, \ldots, n\}), \tag{1}$$

where *uid* is the unique identifier of the user , (x_i, y_i) represents the spatial coordinates at time t_i, and n depicts the trajectory length.

Definition 2. (**Individual Trajectory Prediction**): Given the past M records T^t_{t-M+1} of a trajectory T, and the time information D^{t+N}_{t+1} for the future N records, the objective of individual trajectory prediction is to determine a mapping function $f(\cdot,\cdot)$ with parameters θ, which predicts the locations of the subsequent N records:

$$f(T^t_{t-M+1}, D^{t+N}_{t+1}) = \{(x_j, y_j) \mid j = t+1, \ldots, t+N\}, \tag{2}$$

where

$$T^t_{t-M+1} = (uid, \{(t_i, x_i, y_i) \mid i = t-M+1, \ldots, t\}), \tag{3}$$

and

$$D^{t+N}_{t+1} = \{t_j \mid j = t+1, \ldots, t+N\}. \tag{4}$$

Definition 3. (**Citywide Trajectories**): Citywide trajectories, denoted as $\mathcal{T}^{\mathcal{R}}_{\mathcal{X}}$, refer to a collection of user trajectories within a specified urban area $\mathcal{X}$ and time range $\mathcal{R}$.

Definition 4. (**Citywide Trajectory Prediction**): Given a citywide trajectory dataset for an urban area $\mathcal{X}$ within a past time range $\mathcal{R}_{\text{past}}$, the goal is to predict the future trajectories in the $\mathcal{X}$ over a future time range $\mathcal{R}_{\text{future}}$, using the individual trajectory prediction function $f(\cdot,\cdot)$ with parameters θ:

$$\mathcal{T}^{\mathcal{R}_{\text{past}}}_{\mathcal{X}}, \mathcal{R}_{\text{future}} \xrightarrow[\theta]{f(\cdot,\cdot)} \mathcal{T}^{\mathcal{R}_{\text{future}}}_{\mathcal{X}}. \tag{5}$$

3 Methodology

3.1 Problem Reformulation

It is widely known that LLMs imply vast amounts of knowledge, including the common understanding of human mobility behaviors and patterns [7, 8]. To explore the potential of LLMs in human mobility modeling, we re-framed the trajectory prediction (Equation 2) as a Q&A task with instructions (Figure 1):

- The **instruction block** offers guidelines on the model's *#role*, *#target environment*, *#definition and example of trajectory*, *#task description*, and the *#format of the output.*
- The **question block** includes the user's *#historical trajectory* and the *#time information* for which location predictions are required.
- While the **answer block** is the *#predicted future trajectory* in JSON format, which we expect as the model's response given the previous two blocks.

[1]https://wp.nyu.edu/humobchallenge2024/

Instruction

[Role] You are a helpful assistant that predicts human mobility trajectories in a city.
[Introduction #*Environment*] The target city is divided into equally sized cells, creating a 200 x 200 grid. We use coordinate **<x>,<y>** to indicate the location of a cell within the target area. The horizontal coordinate **<x>** increases from left to right, and the vertical coordinate **<y>** increases from top to bottom. The coordinates of the top-left corner are (0, 0), and the coordinates of the bottom-right corner are (199, 199).
[Introduction #*Trajectory Definition*] A trajectory is a sequence of quadruples ordered by time. Each quadruple follows the format **<day_id>, <timeslot_id>, <x>, <y>**. It represents a person's location **<x>, <y>** at the timeslot **<timeslot_id>** of day **<day_id>**. The **<day_id>** is the index of day, representing a specific day. Each day's 24 hours are discretized into 48 time slots with a time interval of 30 minutes. **< timeslot_id>** is the index of the time slot, ranging from 0 to 47, representing a specific half-hour in a day.
[Introduction #*Trajectory Example*] Let me give you an example of a quadruple to better illustrate what is a record in a trajectory. For instance, a sequence (1,12,124,121) indicates that an individual was located at cell 124,121 between 11:30 and 12:00 on day 1.
[Task Description] You will receive an individual's trajectory in the target city, with some cell coordinates **<x>,<y>** that were missed and marked as 999,999. Please replace all instances of 999 with your predictions.
[Format emphasizing] Please organize your answer in Json object containing following keys: *{"prediction": here should be the missing part of sequence only, without adding any extra things.}* Do not write any code, just inference by yourself; do not provide any other things in your response besides the Json object.

Question

Here is the data I wish you to predict:
day_id, timeslot_id, x, y
<0,0,199,199>,<0,1,198,196>, <0,2,195,196>, …<59,47,198,196>,
<60,1,**999,999**>,<60,2,**999,999**>,<61,1,**999,999**>,…<74,47,**999,999**>.

Answer

{ "prediction":
[[60,1,**92,94**], [60,2,**93,95**], [61,1,**93,94**],... [74,47,**101,108**]]
}

Figure 1: Framing human mobility prediction as a Q&A task with instruction.

Compared to the **question block only** approach, this design offers LLMs a more comprehensive task context to aid in reasoning.

⋆ **Dataset adaptation**. It is important to note that the framework in Figure 1 has been specifically adapted for the Human Mobility Challenge 2024 dataset [10].

3.2 Instruction Tuning

Limitations of zero-shot approach. We conducted exploratory experiments with GPT-3.5 and 4, and found that our Q&A framework is capable of producing decent results in a zero-shot setting. However, due to the high cost of API calls and the potential risk of data leakage in production environments, open-source models may be a better option. Nevertheless, further exploration revealed that for open-source models (i.e., the Llama series), zero-shot is not a viable solution—the models struggled to yield outputs that adhered to the desired format or aligned with the task objectives. Even when producing seemingly acceptable results occasionally, the prediction accuracy was still poor.

Table 1: Basic statistics of dataset.

Attribute	City			
	A	*B*	*C*	*D*
Time Coverage	75 days			
Time Granularity	30 minutes (48 time slots per day)			
Spatial Granularity	500 meters × 500 meters (200 × 200 grids)			
# Trajectories	100,000	22,000	17,000	3,000
Avg. Traj. Len.	1115.35	998.91	948.08	1450.83

Fine-tuning. To overcome challenges faced in zero-shot scenarios, we fine-tuned Llama-3-8B [9] through *instruction tuning* [11]: (1) *Data preparation*: we sampled a portion of users from the training set and constructed fine-tuning corpus with Figure 1 as the template, where the **instruction block** and **question block** served as the model input, and the **answer block** as the expected output. (2) *Parameter-efficient fine-tuning*: given a large number of parameters in LLMs, we applied Low Rank Adaptation (LoRA) adapters [12] to optimize fine-tuning efficiency, only targeting the key modules in the model, such as the query, key, value, and output projections of transformers, as well as the gate, up, and down projections. (3) *Loss function*: we utilized token-level cross-entropy as the loss function directly, converting the complex spatial-temporal learning problem as a Q&A fine-tuning task in the NLP domain.

4 Experiments

4.1 Experimental Setup

Dataset. The dataset we used comes from the Human Mobility Challenge 2024 [10], which includes human mobility data from four metropolitan areas in Japan (cities A, B, C, and D) over a 75-day period. Each trajectory is represented as a quadruplet <x, y, day, time slot>, where x and y are discretized spatial coordinates mapped onto a 200 × 200 grid, and day and time slots represent discretized temporal information with half-hour intervals. The dataset contains complete trajectories for 100,000 individuals from city A and for 25,000, 20,000, and 6,000 individuals in cities B, C, and D respectively. For cities B, C, and D, the location data for the last 3,000 individuals during days 61 to 75 is masked; we excluded this subset of users directly in this study. Table 1 shows the basic statistics of the dataset. The goal is to infer the future trajectories of cities B, C, and D between days 61-75 using the past trajectories from days 1-60, corresponding to parameters $\mathcal{R}_{past}$ and $\mathcal{R}_{future}$ as defined in *Definition 4.* To evaluate the model performance, for each city, the data is further split into training and evaluation sets in an 8:2 ratio by user ids. Moreover, given the efficiency constraints of LLMs in inference tasks—where predicting a single user's trajectory takes approximately 5 minutes—we randomly selected 100 users from the validation set to form a smaller evaluation subset for performance analysis within a restricted timeframe.

Baselines. We used the **champion** model [13] of the Human Mobility Prediction Challenge 2023, namely **LP-Bert**, as our baseline. This model treats the trajectory prediction task as a BERT-based imputation task, where each location record's full information is represented as a token, with a spatial-temporal embedding layer converting it into token embedding. Since the model does not support cross-city predictions, we modified the output layer by adding a city embedding module to enable the multi-city scenario.

Table 2: Model performance comparison.

Model	Average DTW (↓)			Average GEO-BLEU (↑)			Mean Rank
	B	*C*	*D*	*B*	*C*	*D*	
LP-Bert [13]	**23.30**	23.81	38.89	0.3093	0.2682	0.3033	4.17
Llama-3-8B-Mob *w/ B*	26.32	22.49	<u>34.41</u>	0.3322	<u>0.2895</u>	<u>0.3157</u>	2.50
Llama-3-8B-Mob *w/ C*	31.58	23.75	34.49	<u>0.3399</u>	0.2891	0.2833	3.67
Llama-3-8B-Mob *w/ D*	28.75	<u>22.20</u>	38.46	0.3251	0.2765	0.3056	3.50
Llama-3-8B-Mob *w/ A+B*	<u>25.39</u>	**20.57**	**31.94**	**0.3541**	**0.2969**	**0.3217**	**1.17**

Parameter Settings. Our base model is an optimized version of Llama-3-8B, which is 2-5x faster than the original. It uses 4-bit quantization to reduce GPU memory usage by 70%. Readers can access it here[2]. During fine-tuning, the rank of the LoRA adapters was set to 16. The training batch size was 1, with gradient accumulation over 4 steps and a learning rate of 2e-3. The number of fine-tuning epochs was set to 3. AdamW optimizer was used with a cosine learning rate scheduler and a weight decay of 0.01.

Environment. Experiments were conducted on a linux server equipped with 2 Intel(R) Xeon(R) Silver 4310 CPUs and 4 NVIDIA RTX A6000 GPUs, each offering 48 GB of memory. All training and inference procedures were tracked and logged using the Wandb tool[3], ensuring reproducibility and consistent reporting.

Evaluation Metrics. We used Dynamic Time Warping (DTW) [14] and GEO-BLEU [15] to evaluate the quality of predictions. DTW measures the shape similarity between predicted and ground truth trajectories, while GEO-BLEU, a geospatial variant of BLEU, incorporates spatial proximity into n-gram matching to assess the similarity of geospatial sequences.

4.2 Effectiveness Evaluation

In this section, we assess the effectiveness of LP-Bert and Llama-3-8B-Mob. LP-Bert was trained using the complete training data of cities A, B, C, and D. In contrast, training Llama-3-8B-Mob with the entire dataset would be impractical and time-consuming, potentially taking several months (see Section 4.3). Therefore, we conducted instruct tuning separately with the training data from cities B, C, and D, resulting in three distinct versions of Llama-3-8B-Mob. Table 2 shows the performance of these models on the validation set. Notably, even with simple fine-tuning based solely on single city data, our model (all three variants) significantly outperforms last year's leading model which was trained on data from all cities. We attribute this superiority to Llama-3-8B's inherent understanding of human mobility patterns as well as its strong reasoning ability. Moreover, Llama-3-8B-Mob *w/ single city* performs remarkably well in predictions for other cities, showcasing its generalization ability across different urban environments without the need for city-specific data! Furthermore, it can be found that fine-tuning with City B yields the best results (mean rank = 2.5). However, the average trajectory length of City B is 998.91, which may not encompass some long trajectory scenarios. Therefore, we randomly selected 1,000 trajectories from A (average trajectory length to 1115.35) to fine-tune together with B, resulting in improved outcomes: all metrics achieved SOTA results except for DTW@City-B.

Due to limitations in computational resources and time, we only investigated four data fine-tuning settings here (B, C, D, A+B). Even

[2]https://huggingface.co/unsloth/llama-3-8b-bnb-4bit
[3]https://github.com/wandb/wandb

Table 3: Training and inference efficiency.

Model	# Trainable Parameters	Training		Inference	
		GPU Usage	t_{total}	*GPU Usage*	t_{infer}
LP-Bert [13]	12.20 M	25.97 GiB	2.77 d	1.49 GiB	13.94 ms
Llama3-8B-Mob *w/ A+B*	41.94 M	43.11 GiB	6.64 d	14.86 GiB	225.61 s

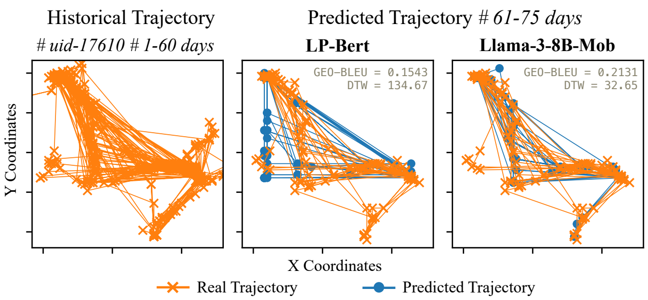

Figure 2: Visualization comparison of prediction results.

though, we believe that there exists an optimal data combination that covers more scenarios and corner cases while minimizing information redundancy.

4.3 Efficiency Evaluation

Although Llama-3-8B-Mob demonstrates strong predictive capabilities, we must acknowledge that its training and inference costs are significantly higher compared to LP-Bert. Table 3 provides an overview of the training and inference efficiency of both models. As shown, the training time for Llama-3-8B-Mob *w/ A+B* reaches 6.64 days, which is 2.4x longer than that of LP-Bert. Additionally, the average inference time per trajectory (batch size = 1) is 225.61 seconds—**16,000x** slower than LP-Bert. Due to the auto-regressive nature of the Llama-3-8B, the inference time increases linearly with the length of the trajectory, with the longest inference time in the validation set reaching up to 15 minutes. This presents a significant challenge when applying LLMs to long-term trajectory prediction.

4.4 Case Study

We conducted a case study on a randomly selected individual from city B, as shown in Figure 2. The left panel displays the user's historical trajectory for days 1-60, while the middle and right panels show the predictions from LP-Bert and Llama-3-8B, respectively. As observed, LP-Bert tends to consistently predict trajectories with right-angled triangular shapes, which clearly deviate from general human mobility patterns. In contrast, Llama-3-8B accurately replicates an individual's movement behavior—the orange and blue lines nearly overlap. Notably, LP-Bert's behavior is not an isolated case; in many other examples, it also tends to predict regular geometric shapes (e.g., squares). Additionally, the gray text shows the DTW and GEO-BLEU scores for both models, indicating that Llama-3-8B-Mob significantly outperforms LP-Bert.

5 Conclusion

This study proposes an LLMs instruction tuning approach for long-term human mobility prediction across multiple cities. We validated the effectiveness and the superiority of the proposed method using large-scale human mobility data from four metropolitan areas in Japan. In the future, we will improve our work from three aspects: (1) Recognizing that the effectiveness of fine-tuning LLMs is highly dependent on data quality, *designing better data selection strategies* (currently random) towards more efficient and effective fine-tuning will be next focus. (2) The slow inference time greatly affects the practicality of the Llama-3-8B-Mob. We will continue to explore the latest advances in *efficient inference techniques* and apply them to our model. (3) Expanding validation to *more trajectory datasets*, beyond human trajectories and grid-level data.

Acknowledgments

This work was supported by JST SPRING JPMJSP2108, JSPS KAKENHI JP24K02996, and JST CREST JPMJCR21M2 including AIP challenge program, Japan.

References

[1] Renhe Jiang, Xuan Song, Zipei Fan, Tianqi Xia, Quanjun Chen, Satoshi Miyazawa, and Ryosuke Shibasaki. Deepurbanmomentum: An online deep-learning system for short-term urban mobility prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[2] Renhe Jiang, Zhaonan Wang, Zekun Cai, Chuang Yang, Zipei Fan, Tianqi Xia, Go Matsubara, Hiroto Mizuseki, Xuan Song, and Ryosuke Shibasaki. Countrywide origin-destination matrix prediction and its application for covid-19. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part IV 21*, pages 319–334. Springer, 2021.

[3] Xiaohang Xu, Toyotaro Suzumura, Jiawei Yong, Masatoshi Hanai, Chuang Yang, Hiroki Kanezashi, Renhe Jiang, and Shintaro Fukushima. Revisiting mobility modeling with graph: A graph transformer model for next point-of-interest recommendation. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pages 1–10, 2023.

[4] Jie Feng, Yong Li, Chao Zhang, Funing Sun, Fanchao Meng, Ang Guo, and Depeng Jin. Deepmove: Predicting human mobility with attentional recurrent networks. In *Proceedings of the 2018 world wide web conference*, pages 1459–1468, 2018.

[5] Haochen Zhang, Yuyang Dong, Chuan Xiao, and Masafumi Oyamada. Jellyfish: Instruction-tuning local large language models for data preprocessing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, November 2024. Association for Computational Linguistics.

[6] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large language models are zero-shot time series forecasters. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 19622–19635, 2023.

[7] Xinglei Wang, Meng Fang, Zichao Zeng, and Tao Cheng. Where would i go next? large language models as human mobility predictors. *arXiv preprint arXiv:2308.15197*, 2023.

[8] Peibo Li, Maarten de Rijke, Hao Xue, Shuang Ao, Yang Song, and Flora D Salim. Large language models for next point-of-interest recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1463–1472, 2024.

[9] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[10] Takahiro Yabe, Kota Tsubouchi, Toru Shimizu, Yoshihide Sekimoto, Kaoru Sezaki, Esteban Moro, and Alex Pentland. Yjmob100k: City-scale and longitudinal dataset of anonymized human mobility trajectories. *Scientific Data*, 11(1):397, 2024.

[11] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[13] Haru Terashima, Naoki Tamura, Kazuyuki Shoji, Shin Katayama, Kenta Urano, Takuro Yonezawa, and Nobuo Kawaguchi. Human mobility prediction challenge: Next location prediction using spatiotemporal bert. In *Proc. of the 1st International Workshop on the Human Mobility Prediction Challenge*, pages 1–6, 2023.

[14] Fumitada Itakura. Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on acoustics, speech, and signal processing*, 23(1):67–72, 1975.

[15] Toru Shimizu, Kota Tsubouchi, and Takahiro Yabe. Geo-bleu: similarity measure for geospatial sequences. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, pages 1–4, 2022.