Machine learning discovery of regional and social disparities in electric vehicle charging reliability with GPT5

Yifan Liu, Lindsey Snyder, Omar I. Asensio
Georgia Institute of Technology

Georgia Tech • Ivan Allen College of Liberal Arts
Jimmy and Rosalynn Carter School of Public Policy

Why EV charging reliability matters?

- **Transportation** is the second-largest source of global emissions and the top source in many developed countries
- Electrifying vehicles is critical to **decarbonization**, and expanding reliable **charging infrastructure** is essential for EV adoption (Li et al., 2020).
- However, **charging failures** have been documented across the U.S., Europe, and Asia (Rempel et al., 2024, Liu et al, 2023; Asensio et al., 2020).
- **Machine learning** has emerged as a key strategy for charging management, including algorithm-based decision-making, load balancing, and demand forecasting (Yaghoubi et al., 2024).

Challenges for ML/AI discovery

- Prior methods for assessing charger reliability, most of which rely on citizen-generated data and expensive expert annotations, lack the detection **accuracy** required for **large-scale inference**.
- Current methods also fail to capture **regional and social disparities** in consumer-reported reliability.

Our approach to measuring EV charging **reliability** and **disparity**

- Review level: develops a **zero- and few-shot chain-of-thought learning pipeline** to detect charging reliability from 838,785 consumer reviews; integrates expert-guided prompt refinement to reduce Type I and II errors, achieving high **detection accuracy (F1 = 0.97)**.
- Station level: reliability score represents the share of reviews without reliability issues.
- County level: combines reliability detection with **geographic disparity indices** (i.e., Shannon Evenness Index) to measure intra-county variation in reliability.

$$E_c = \frac{-\sum_{k=1}^{m} p_k \ln p_k}{\ln m}$$

AI alignment: expert prompts with GPT4/5 outperform popular models such as ClimateBERT

Table 1: Model performance and efficiency for detecting charging reliability issues in user reviews

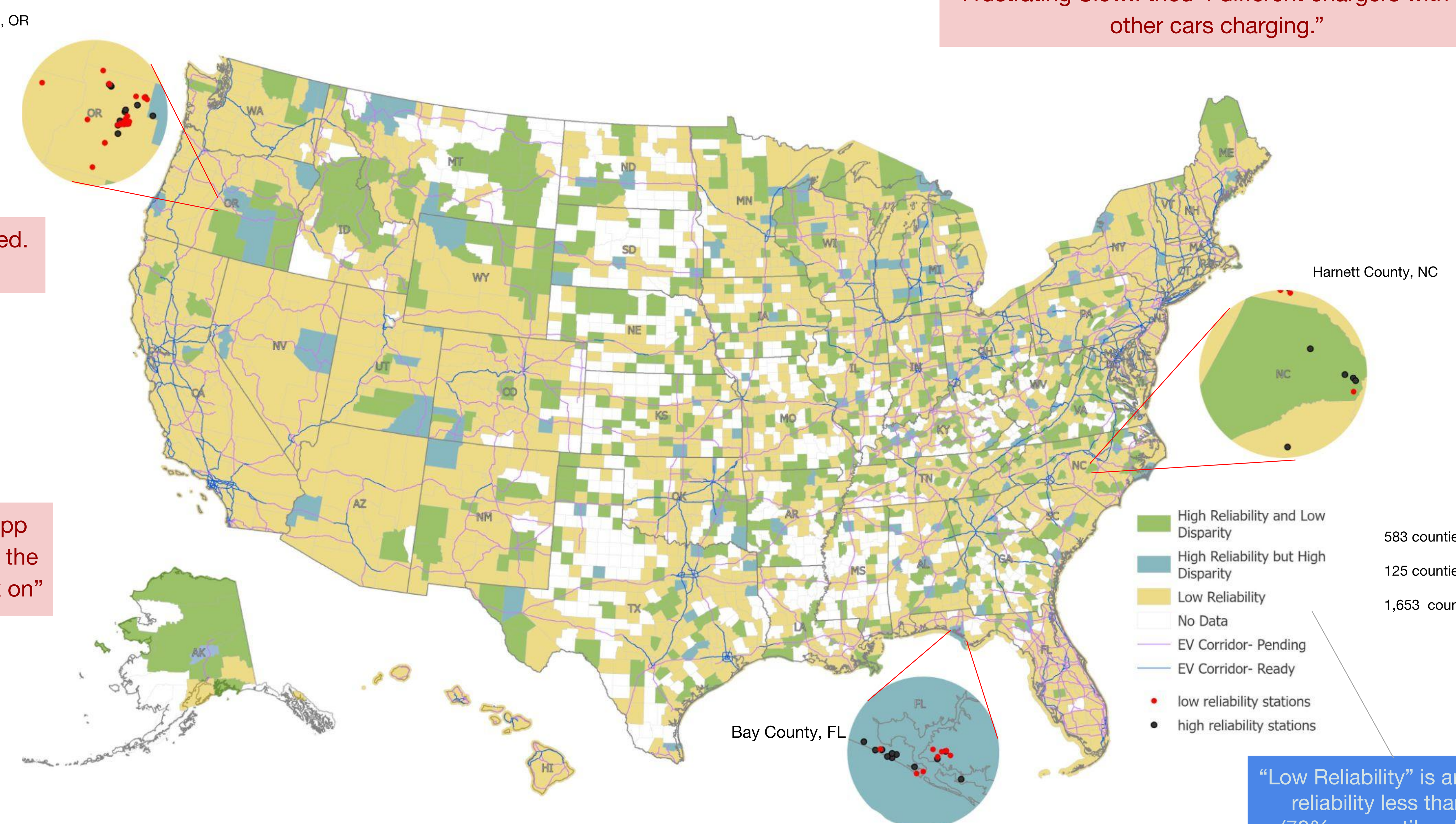
Model	Input	Accuracy % (s.d.)	F1 Score (s.d.)	Training cost	Type I error rate	Type II error rate
Few/zero-shot models with incrementally optimized CoT prompts						
GPT-5 few-shot [gpt-5-2025-08-07]	CoT prompt: instruction + definitions + examples + counterexamples	97.10% (1.66)	0.96 (0.024)	0	2.1%	3.4%
GPT-4 few-shot [gpt-4-turbo-2024-04-09]	CoT prompt: instruction + definitions + examples + counterexamples	97.90% (1.60)	0.97 (0.018)	0	1.1% (↑ 0.4%)	3.4% (↓ 5.0%)
GPT-4 few-shot [gpt-4-turbo-2024-04-09]	CoT prompt: instruction + definitions + examples	95.90% (1.45)	0.95 (0.016)	0	0.7% (↓ 0.2%)	8.4% (↓ 7.2%)
GPT-4 zero-shot [gpt-4-turbo-2024-04-09]	CoT prompt: instruction + definitions	92.60% (3.17)	0.91 (0.037)	0	0.9% (↓ 20.3%)	15.6% (↑ 6.1%)
GPT-4 zero-shot [gpt-4-turbo-2024-04-09]	Prompt: instruction	90.80% (2.44)	0.89 (0.034)	0	21.2% (init.)	9.5% (init.)
Fine-tuned models (with significantly higher labeling costs) for reference						
ClimateBERT fine-tuned	Training set: 4,000	90.90% (1.18)	0.90 (0.012)	$$$ expert labeling	10.7%	7.3%
GPT-4o fine-tuned [gpt-4o-2024-08-06]	Training set: 4,000	97.30% (2.11)	0.97 (0.027)	$$$ expert labeling	2.4%	3.0%

* The accuracies and F1 scores are computed for the reliability_issue label using the same test set of 1,000 observations, with ground truth determined by expert human votes. We assume 25M input and 1M output tokens for 1M reviews for the estimation of inference costs in the first model.

“Frustrating Slow.. tried 4 different chargers with no other cars charging.”

“All three chargers are damaged. Car does not charge.”

“Downloaded the new app and tried to charge using the app, but screen got stuck on”



U.S. county classification by EV charging reliability and disparity (2012–2024)

“Low Reliability” is an average reliability less than 0.80 (70% percentile reliability)

“Low Disparity” is less than 0.40, meaning a lower diversity of reliability scores in the county.

Takeaways

- **High accuracy, low cost**
  - Few-shot CoT model hits 97.9% accuracy (F1 = 0.97), outperforming prior work at a fraction of the cost.
  - Expert-guided prompt refinement cuts false positives by 94.9% and false negatives by 64.3%.
- **Widespread reliability gaps nationwide**
  - 1,650+ counties show low reliability, affecting 300 million residents, concentrated in urban hubs and EV corridors.
- **Charging disparity and policy implications**
  - “Charging deserts” expose inequities, causing inconsistent charging experience
  - Calling for performance-based EV charging policies.
- **Next steps:** Integration into causal inference and prediction studies in electric mobility.

References:

1. Asensio, O. I., Alvarez, K., Dror, A., Wenzel, E., Hollauer, C., & Ha, S. (2020). Real-time data from mobile platforms to evaluate sustainable transportation infrastructure. *Nature Sustainability*, *3*(6), 463-471.
2. Li, S., Tong, L., Xing, J., & Zhou, Y. (2017). The market for electric vehicles: indirect network effects and policy design. *Journal of the Association of Environmental and Resource Economists*, *4*(1), 89-133.
3. Liu, Y., Francis, A., Hollauer, C., Lawson, M. C., Shaikh, O., Cotsman, A., ... & Asensio, O. I. (2023). Reliability of electric vehicle charging infrastructure: A cross-lingual deep learning approach. *Communications in Transportation Research*, *3*, 100095.
4. Rempel, D., Cullen, C., Matteson Bryan, M., & Vianna Cezar, G. (2024). Reliability of open public electric vehicle direct current fast chargers. *Human Factors*, *66*(11), 2528-2538.
5. Yaghoubi, E., Khamees, A., Razmi, D., & Lu, T. (2024). A systematic review and meta-analysis of machine learning, deep learning, and ensemble learning approaches in predicting EV charging behavior. *Engineering Applications of Artificial Intelligence*, *135*, 108789.

Acknowledgements: We gratefully acknowledge funding support by the the National Science Foundation (Grant Nos. 1931980 and 1945332), Microsoft Azure Research, and the Georgia Tech Energy Policy and Innovation Center (EPIcenter).