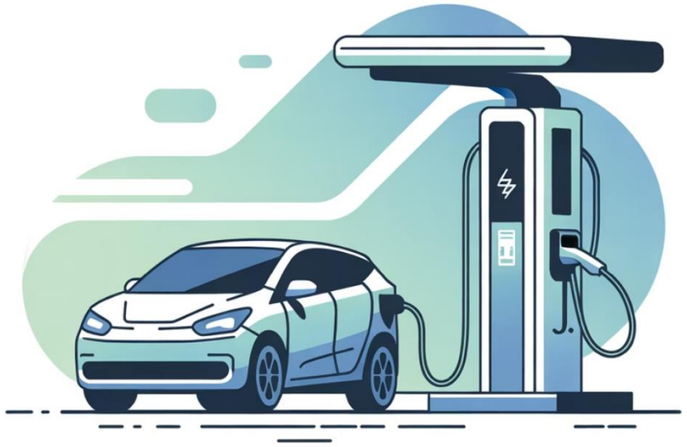


Machine Learning Discovery of Regional and Social Disparities in Electric Vehicle Charging Reliability

Yifan Liu, Lindsey Snyder, Omar I. Asensio

Georgia Institute of Technology

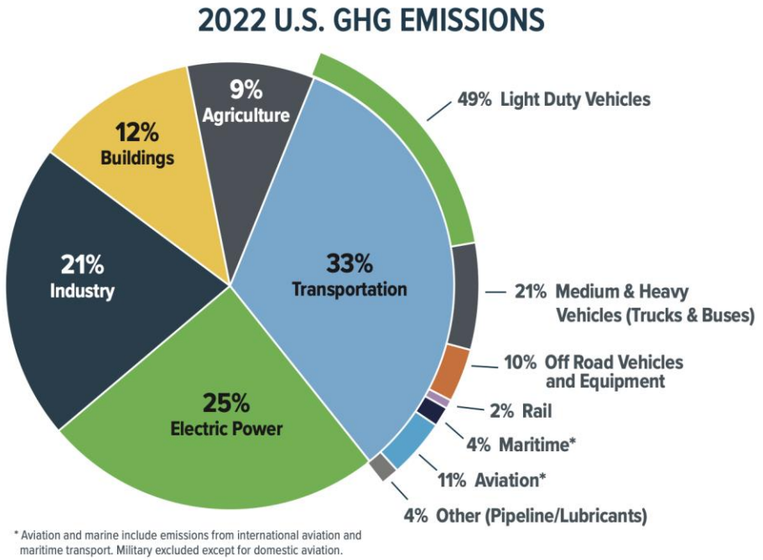
Georgia Tech • Ivan Allen College of Liberal Arts
Jimmy and Rosalynn Carter School of Public Policy

NSF

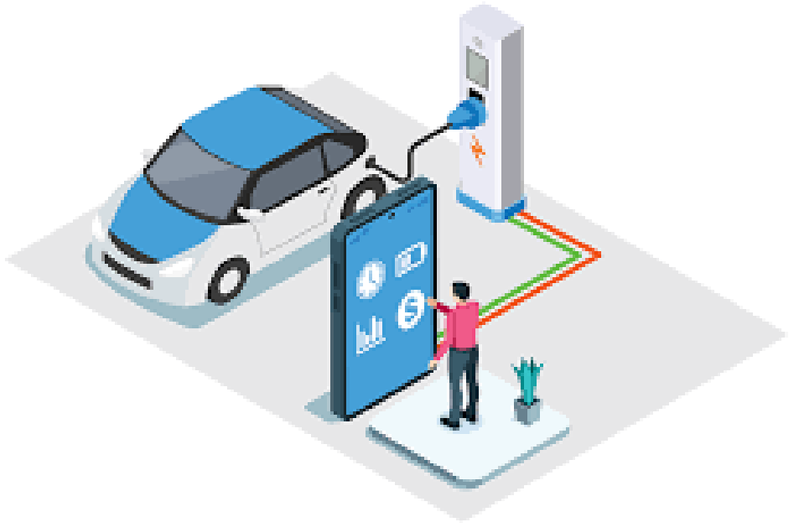
Microsoft Azure

NEURAL INFORMATION PROCESSING SYSTEMS

Why EV charging reliability matters



Source: U.S. Department of Transportation (2024)



> *“Frustrating. Slow. Tried 4 different chargers with no other cars charging.”* - Deschutes County, OR

> *"All three chargers are damaged. Car does not charge.”* - Harnett County, NC

Charging infrastructure is more cost-effective in promoting EV adoption (Li et al., 2017; Springel et al., 2021)

Challenges for AI/ML discovery

- Current methods (Asensio et al., 2020; Yu et al., 2025) for assessing reliability at a large scale:
  - expensive expert labelling and unbalanced classes in the dataset;
  - lack the accuracy needed for large-scale inference;
  - fail to capture regional and social disparities in consumer-reported experiences.

Our approach: measuring 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 feedback into an iterative error-analysis loop to optimize prompts (Wei et al., 2022; Kim et al., 2025) and systematically reduce Type I and Type II errors, resulting in **high detection performance (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.

$$\bar{R}_c = \frac{1}{S_c} \sum_{i \in c} \left(\frac{1}{T_i} \sum_{t \in T_i} R_{i,t} \right)$$

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

Model performance:
incrementally optimized prompts outperform ClimateBERT baseline models

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)	$$$ expert label-ing	10.7%	7.3%
GPT-4o fine-tuned `[gpt-4o-2024-08-06]`	Training set: 4,000	97.30% (2.11)	0.97 (0.027)	$$$ expert label-ing	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.

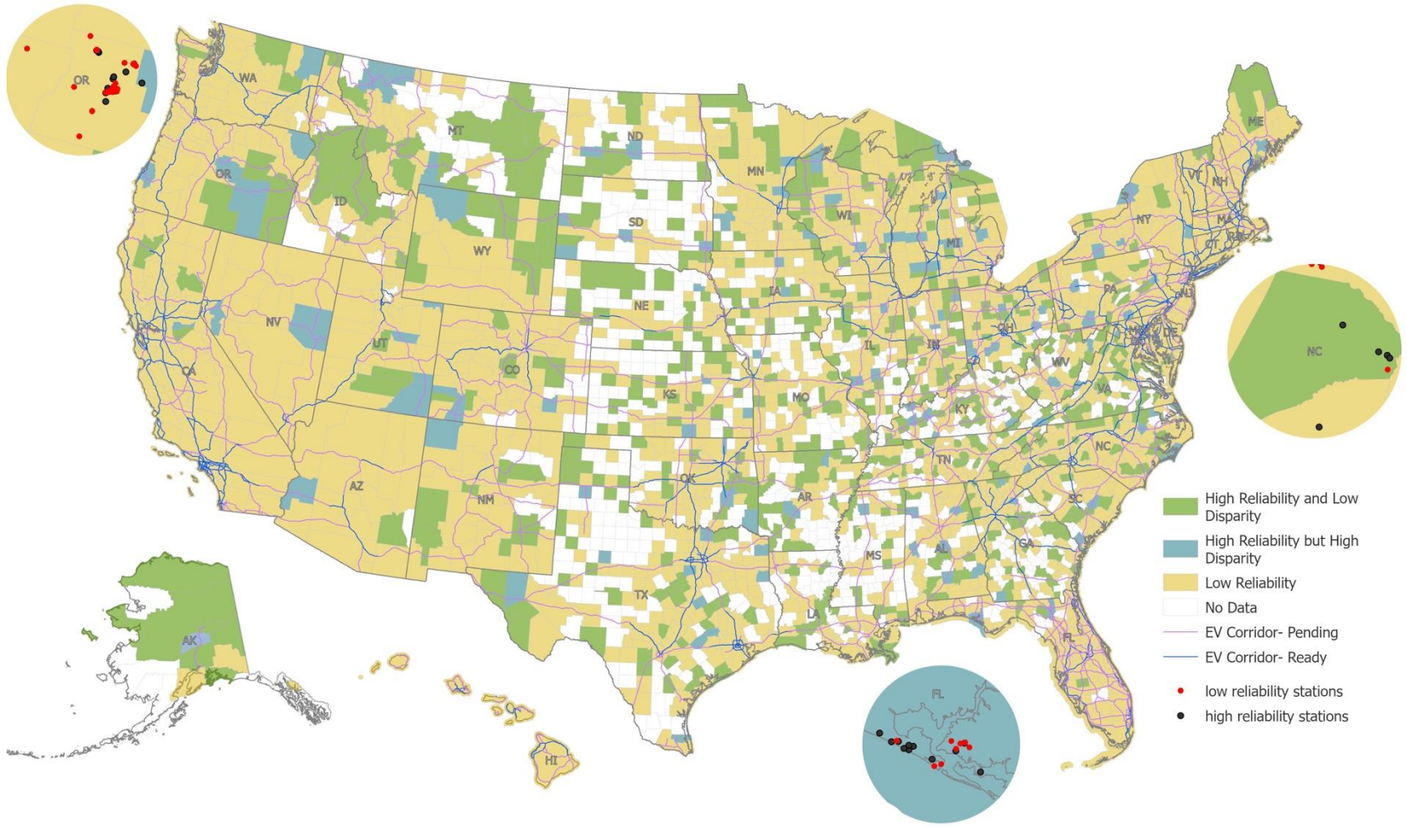
Model Performance

97.9% accuracy achieved through few-shot chain-of-thought learning

94.9% reduction in false positives through expert refinement

64.3% reduction in false negatives compared to baseline approaches

Widespread charging reliability issues across U.S. counties



1,653 Counties

Low Reliability: average reliability below 0.80 (70th percentile threshold)

125 Counties

High Reliability but High Disparity: significant variation in charging experiences within county boundaries (Shannon Evenness Index above 0.4)

583 Counties

High Reliability and Low Disparity: no significant issues in charging reliability issues

Key takeaways and next steps

1

High Accuracy, Low Cost

Few-shot chain-of-thought model achieves 97.9% accuracy (F1 = 0.97), dramatically outperforming prior work at a fraction of the cost.
Expert-guided refinement cuts false positives by 94.9% and false negatives by 64.3%.

2

Nationwide Reliability Gaps

Over 1,650 counties show low reliability, affecting approximately 300 million residents. These gaps are concentrated in urban hubs and major EV corridors where charging demand is highest.

3

Charging Deserts and Equity

"Charging deserts" expose critical infrastructure inequities, causing wildly inconsistent charging experiences within the same geographic area. These disparities call for performance-based EV charging policies.

Future Directions

- Integration into **causal inference studies** examining factors driving charging reliability
- Development of **predictive models** for charging infrastructure planning
- Application to **policy evaluation** and performance-based incentive design