**Slide 1:**

Good morning, everyone. My name is Chenyi Xiang, and today I'll be presenting on **Generative Modeling for Synthetic Data**. This presentation is part of my internship project with GEICO's AI & ML Core Technologies Team.

**Slide:**

A quick introduction about me: I hold a bachelor's degree in applied linguistics and a master's degree in computer science from Northeastern University. Although it may seem that I changed my major, programming languages are also a type of language and my background in linguistics helped me develop a better understanding of programming languages and how they are constructed.

Throughout my academic and professional journey, I've been passionate about leveraging machine learning to create innovative solutions, particularly those with practical applications and real-world impact. I enjoy exploring new technologies and pushing the boundaries of their real-world use.

A fun fact about me is that I can solve a Rubik's Cube in under 20 seconds.

**Slide 2: Business Value**

**GEICO requires data to perform functional and load testing for our telematics production pipelines.**

Telematics uses data from sensors and devices in vehicles to monitor driving behavior, vehicle status, and more. For a company like GEICO, having robust telematics data is crucial for testing various aspects of our systems:

* **Functional Testing** ensures that all components and systems within the telematics pipeline work as intended. For example, verifying that data from vehicle sensors is accurately captured, transmitted, and processed.
* **while Load Testing** involves testing the telematics systems under heavy data loads to ensure they can handle peak conditions. It's crucial to simulate high traffic scenarios to ensure system reliability and performance.

**However, obtaining production data poses several challenges such as privacy and security concerns, data quality issues, costs, and regional expansion limitations.**

Let's break down these challenges:

* **Privacy and Security Concerns:**
  + **Privacy:** Using actual data for testing could risk exposing the private information of customers.
  + **Security:** Storing and handling real data necessitates strict security measures to protect against data breaches
* **Quality Issues:**
  + Real-world data can be inconsistent, with gaps or errors that could skew testing results. These quality issues might arise from such as hardware malfunctions, data corruption, making it difficult to ensure comprehensive testing.
* **Costs:**
  + Collecting and storing real data or purchasing high quality data from commercial sources is expensive.
* **Regional Expansion Limitations:**
  + Real-world data is often region-specific, which means that expanding telematics services to new regions or countries requires collecting new datasets.

Therefore, synthetic data is crucial as it provides a viable alternative that ensures data availability without compromising privacy or incurring high costs.

**Slide 3: Telematics Data**

**Our telematics data here at GEICO consists of time-series trip information. This data has been categorized into the following three tiers to make it easier to understand our project goals. The higher-level data in tier 1 will be the focus of my presentation.** And below is the **existing data synthesis pipeline** we use to generate these 3 data tiers.

**Slide 4: Architecture**

On this slide, you can see the architectural comparison between the existing pipeline and the new pipeline. Our new pipeline leverages a NanoGPT model to produce second-by-second coordinates directly from input trip parameters, which adds more realism and variability compared to the traditional method. To be specific, the components involving loading map data and interpolating latitude/longitude points are replaced by our new Nano GPT model.

**Slide 5: GPT Model Overview**

So, what is GPT? It's a transformer-based architecture traditionally used for generating text by predicting future words from a starting phrase or word. In our case, we've adapted GPT to predict nodes and edges of a graph, effectively generating realistic coordinates and sensor data from a given starting point. This innovative approach helps us generate synthetic data that is not only realistic but also diverse, with natural variability built in.

**Slide 6: Metrics**

To measure the effectiveness of our generated data, we focus on two key metrics:

* **Smoothness:** ensuring that the generated data contains optimal paths learned from the training dataset and does not take a roundabout path to the destination node.
* **and Groundedness:** verifying that the generated data matches real-world patterns and is consistent with the original map that our model is trained on.

Note that our results are based on a sample size of 20000 nodes over several generated trips from different starting points

**Slide 7: Smoothness**

The smoothness metric is defined as the ratio of the shortest path distance in the original map to the total distance of the model-generated route. This measures whether our model is generating routes that are not only realistic but also optimal.

Our model achieves a smoothness ratio of 99.4%, indicating that the generated routes are nearly identical compared to real-world paths in our training dataset.

**Slide 8: Groundedness**

For groundedness, we divide it into two dimensions: validity and **Connectivity.**

Validity is defined as the proportion of nodes generated by the model that are present in the original map data – this measures the tendency of our model to hallucinate connections or roads that don’t exist.

And connectivity is defined as the proportion of valid nodes that are connected via edges in the original map data, which measures whether our model can generate nodes in a valid sequence relative to each other.

* **Validity:** 89.6% of the generated nodes are present in the original map data.
* **Connectivity:** 95% of these valid nodes are connected via edges, demonstrating that our model-generated data aligns well with the real-world map data.

**Demo:**

* In this demonstration, I've used our model to generate a trajectory based on the road network. The process involves plotting the road graph and then using the model to generate a sequence of 15 points that form the path. The red line you see on the map represents the trajectory created by the model. This path illustrates the model's capability to understand and navigate the given road network, showcasing its potential applications in route planning and navigation systems. As I run the model, you can see it generates optimal routes from a given starting point, and we can change it as desired to create a separate set of trips.

**Slide 9: Next Steps**

Moving forward, we aim to expand this model to cover additional cities and tiers of data. Specifically, we plan to extend our synthesis pipeline to include more complex data layers, such as those in Tiers 2 and 3, and to replace our regression-based predictions for even more accurate data synthesis.

**Slide 10: Acknowledgement**

I would like to extend my heartfelt thanks to The AI & ML Core Technologies Team for their generous support, especially to my mentors Priyanka and Jingyan as well as my manager Darek. Thank you all for your time and attention. I'm happy to answer any questions you may have.