**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:**

So, here is a intro about me. I hold a Bachelor's degree in Applied Linguistics and a Master's degree in computer science from Northeastern University. Although it seems that I changed my major, but programming language is also a language.

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

As a fun fact about me, I used to be able to solve a Rubik's Cube in under 20 seconds.

**Slide 2: Business Value**

**GEICO requires data to test telematics production pipelines, which is essential for functional and load testing.**

Telematics involves the use of 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 their systems:

* **Functional Testing:** This ensures that all the components and systems within the telematics pipeline work as intended. For example, verifying that data from vehicle sensors is accurately captured, transmitted, and processed.
* **Load Testing:** This 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, 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 this private information.
  + **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 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**

**Telematics data consists of time-series driver trip information. This data is categorized into the following three tiers. The basic data in tier 1 will be the focus of my presentation.**

And below is the **existing data synthesis pipeline** we currently used 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 NanoGPT pipeline. The 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 component of map data loading and the process of interpolate lat/lon points as well as calculating heading is 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. In our case, we've adapted GPT to predict nodes and edges, effectively generating realistic coordinates and sensor data from these nodes. This innovative approach helps us generate synthetic data that is not only realistic but also diverse.

**Slide 6: Metrics**

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

* **Smoothness:** This ensures that the generated data follows consistent and expected rates of motion.
* **Groundedness:** This metric verifies that the generated data matches real-world patterns and is consistent with the original map data.

Before introducing the final metrics results, a precondition that should be noted is that the results are generalized based on a sample size of 20000 nodes to ensure the metric results are accurate.

**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.

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

**Slide 8: Groundedness**

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

They are defined in PPT(read the definition).

we measure two aspects:

* **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 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. (will change the stating point)

**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 leverage 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 mentor Priyanka and Jinyan as well as my manager Darek.

**Slide 11: Questions**

Thank you all for your time and attention. I'm happy to answer any questions you may have.