AI-Assisted Travel Surveys: Enhancing Federal Travel Behavior Data with Generative AI

[1]Nivedya Madankara Kottayi, [2]Daniel Cork, [3]Beth Jarosz, [4]Kabirat Nasiru

[1] Consultant, Association of Public Data Users, [2] Senior Fellow, Massive Data Institute, Georgetown University, [3] Senior Fellow, Massive Data Institute, Georgetown University
[4] Oakridge Institute for Science and Education Data Science Fellow – Research Ambassadors Program through National Center for Science and Engineering Statistics within the National Science Foundation

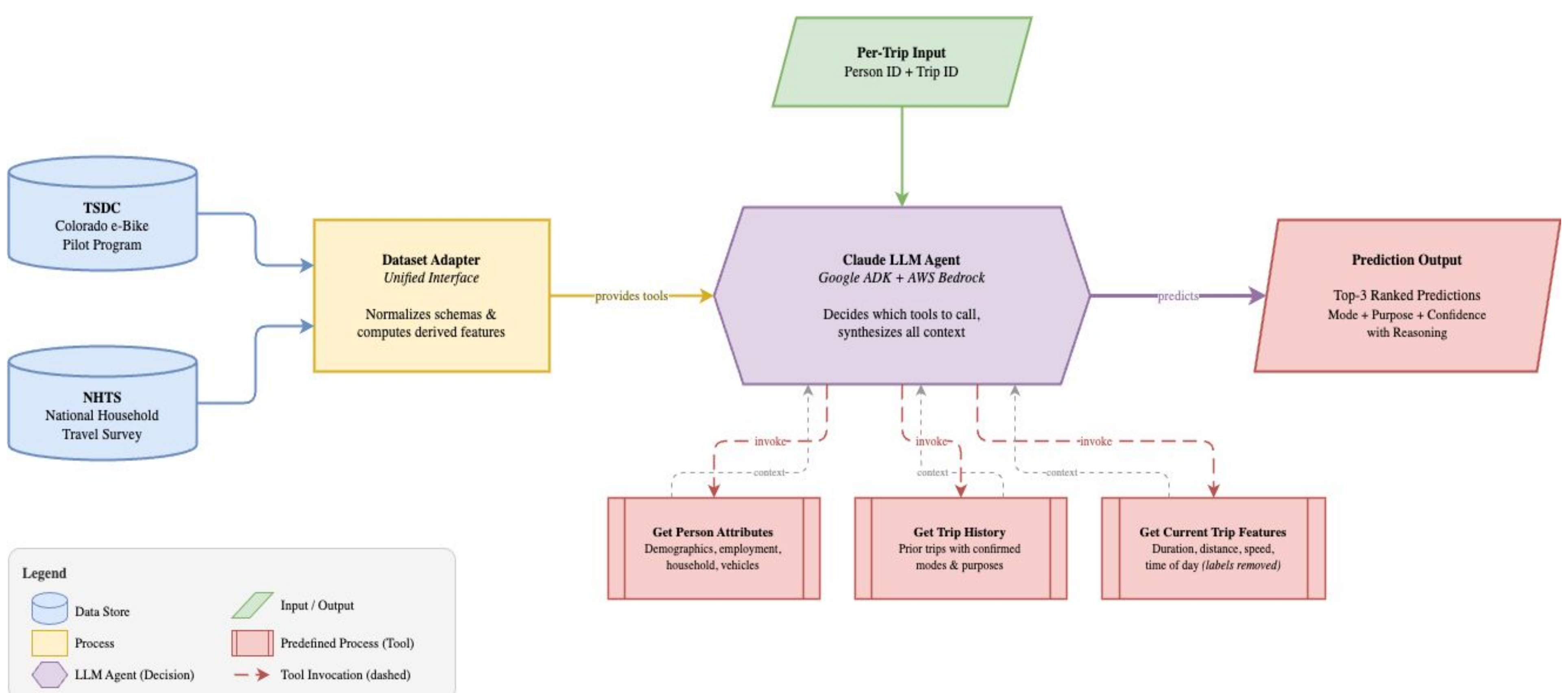
Acknowledgements: This work is supported by the **National Secure Data Service Demonstration (NSDS-D) project.**
This research is being conducted in collaboration with the **Massive Data Institute at Georgetown University.**

Introduction

- Transportation planning relies on detailed travel survey data, including details such as origin, destination, time, mode, and trip purpose.
  - Traditional self-reported methods create **high respondent burden,** leading to fatigue, missing data, and lower data quality.[1]
- This study explores **Generative AI (GenAI)** as a survey-assist tool.
  - Pre-fill trip information and guide respondents using historical travel patterns, reducing manual effort.
- Approach improves survey efficiency, data quality, and consistency, supporting modernized data collection for more reliable transportation planning and policy decisions.
- Supports the modernization of **federal statistical data systems** and the **National Secure Data Service Demonstration (NSDS-D)** by enhancing survey workflows with automated and context-aware inference, improving reliability and scalability of national datasets.

Objective

- Evaluate the effectiveness of GenAI in reducing respondent burden and improving the accuracy and completeness of travel survey data.



Methodology

- Proposed framework uses multiple travel behavior data sources: **National Household Travel Survey (NHTS)** - 2022 NextGen National Household Travel Survey Core Data,[2] and **Transportation Secure Data Center (TSDC)** - 2021–2022 Can Do Colorado E-Bike Full-Scale Pilot Program Study.[3]
  - Differ in travel detail, temporal coverage, and data quality.
- Large Language Model (LLM) based agent was developed using Google ADK and AWS Bedrock (Claude models) to predict transportation mode and trip purpose.
  - Integrate heterogeneous data sources and reason over sequential and contextual information.
- Adapter-based architecture standardizes different datasets into a common structure, enabling consistent access to person demographics, trip history, and trip-level attributes while allowing the model to generalize across datasets.
- Agent analyzes travel behavior using demographic information, historical trip sequences, and current trip features such as distance, duration, speed, and time-of-day (for NHTS), capturing both temporal and contextual patterns.
- Model was evaluated on 200 sampled trips using accuracy, precision, recall, and weighted F1-score.[1]

Performance Measures

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total Predictions}}$$

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Results and Discussion

- Strong performance on **NHTS dataset**.
  - Mode prediction: 91.5% (Top-1) → 99.0% (Top-3).
  - Purpose prediction: 59.5% (Top-1) → 80.5% (Top-3).
- Mode prediction is more accurate than purpose prediction.
  - Purpose prediction may be harder due to ambiguity and overlapping response categories (eg. bike, bike share, pilot e-bike).
- High accuracy driven by dominant vehicle trips.
  - Majority vehicle modes are easily identified using distance and speed.
- Lower performance for minority modes.
- Purpose prediction varies by category.
  - Home and Work → higher accuracy.
  - Meals, social, errands → harder to distinguish.

Table 1: Model generated Top-3 ranked predictions for mode and purpose.

Dataset	Accuracy	Top-1	Top-2	Top-3
NHTS	Mode	91.5	88.5	99.0
NHTS	Purpose	59.5	72.0	80.5
TSDC	Mode	25.1	53.8	69.8
TSDC	Purpose	23.1	40.7	51.3

- Performance is lower when trained on the **TSDC dataset**.
  - Mode: 25.1% (Top-1) → 69.8% (Top-3).
  - Purpose: 23.1% (Top-1) → 51.3% (Top-3).
- Lower accuracy is likely due to dataset complexity.
  - TSCD has more diverse and non-standard. categories, poorer data quality, and missing features (e.g., trip start time).
- Model struggles with similar modes: E-bike, shared ride, and car trips have overlapping patterns.

Table 2: Performance metrics

Dataset	Task	Precision	Recall	F1 Score
NHTS	Mode	0.93	0.92	0.93
NHTS	Purpose	0.61	0.56	0.58
TSDC	Mode	0.33	0.28	0.30
TSDC	Purpose	0.24	0.21	0.22

- Model improvements with advanced LLMs.
  - Newer models (e.g., Claude Opus 4.5) improve performance, especially for purpose prediction and complex datasets.
- Dataset structure strongly impacts performance
  - NHTS: simpler taxonomy → higher accuracy.
  - TSDC: heterogeneous categories → lower accuracy.
- Limited contextual features reduce prediction quality.
  - Missing or sparse inputs significantly affect both mode and purpose inference.

Table 3: Model comparison

Model	NHTS Mode	NHTS Purpose	TSDC Mode	TSDC Purpose
Claude Sonnet 3.7	91.5	59.5	25.1	23.1
Claude Sonnet 4.5	88.5	58.5	30.5	29.9
Claude Opus 4.5	85.4	67.8	42.0	40.5

Conclusion

- **LLM-based agents effectively predict travel mode and purpose**, achieving high accuracy on structured datasets (NHTS), **demonstrating strong potential** for survey automation.
- Performance declines on heterogeneous datasets (TSDC), highlighting the **importance of data quality, standardized taxonomies**, and availability of contextual features.
- Model **validation using precision, recall, and F1-score** confirms the effectiveness of AI-generated trip predictions compared to observed survey responses.
- While advanced models (e.g., Claude Opus) improve accuracy, they require significantly higher **computational cost and tokens**; this study used a small sample (n = 200). However, practical deployment must balance **performance vs. cost and scalability.**
- Findings and lessons learned will be incorporated into a shared resource to inform work in the federal statistical community.

1) Lu, Hannah, Katie Rischpater, and K. Shankari. 2024. "Learning from User Behavior: A Survey-Assist Algorithm for Longitudinal Mobility Data Collection." *Travel Behaviour and Society.* 36 (July): 100761.
2) Federal Highway Administration. (2022). 2022 NextGen National Household Travel Survey Core Data. http://nhts.ornl.gov.
3) Transportation Secure Data Center. (2022). National Renewable Energy Laboratory. Accessed [Oct 23, 2025]. www.nrel.gov/tsdc.