TransMode-LLM: Feature-Informed Natural Language Modeling with Domain-Enhanced Prompting for Travel Behavior Modeling

Meijing Zhang
meijing_zhang@sutd.edu.sg
Singapore University of Technology and Design
Singapore, Singapore, Singapore

Ying Xu*
ying_xu@sutd.edu.sg
Singapore University of Technology and Design
Singapore, Singapore, Singapore

ABSTRACT

Understanding traveler behavior and accurately predicting travel mode choice are at the heart of transportation planning and policy making. Traditional methods relying on raw numbers and structured feature representations have limitations on capturing the complex interdependency and qualitative factors that may impact on travel behavior in the real-world, particularly the rich contextual nuances underlying individual decision-making processes. Large language models (LLMs) with promising capabilities for understanding contextual information across domains provides new pathways for travel behavior modeling. In this study, we propose, TransMode-LLM, an innovative framework designed to predict travel modes from natural language descriptions of travelers and their trips. We start by analyzing the importance of features to identify and select key impacting factors (i.e. individual, household and trip characteristics) to enrich context for decision-making. To enhance the performance of LLMs for transportation-specific tasks, we propose a domain-enhanced prompting strategy that incorporates standardize mode definitions. We further explore various learning paradigms (zero-shot and one/few-shot learning) to understand their impact on travel mode prediction using natural language. Finally, we build an evaluation system to compare the performance of the proposed LLM-based approach against state-of-the-art traditional models. Extensive experiments are conducted on the real-world travel survey dataset and the results demonstrate the competitive performance of LLM-based approach such as prediction accuracy compared to the traditional methods. This study advances the application of LLMs in travel behavior modeling, providing promising and valuable insights for both academic research and transportation policy-making in the future.

KEYWORDS

Large Language Models, Travel Behavior Analysis, Mode Choice Prediction, Transportation Planning

ACM Reference Format:
Meijing Zhang and Ying Xu*. 2025. TransMode-LLM: Feature-Informed Natural Language Modeling with Domain-Enhanced Prompting for Travel Behavior Modeling. In . ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Understanding individual mobility and travel behavior is vitally important in several areas of transportation, including transportation planning, traffic management, and transport policy. The majority of previous studies on travel mode prediction are based on traditional statistical models like multinomial logit models to advanced machine learning techniques, both relying on structured data to establish relationships between traveler characteristics and mode choices [20, 40]. These methods, while valuable, often struggle to capture the complex contextual nuances and heterogeneous patterns that influence travel decisions.

While previous research has compared the effectiveness of various statistical approaches, a fundamental limitation persists: these methods typically operate with predefined variables that may not capture the contextual nuances impacting individual travel decisions. The rapid development of Large Language Models (LLMs) presents an opportunity to address this gap. Trained on vast amounts of text data, LLMs demonstrate significant advantages in possessing remarkable capabilities in understanding context, reasoning about complex relationships, and generating human-like responses.

Recently, researchers have shown an increased interest in applied the advanced techniques from LLMs to understand and predict human behavior across various domains including law, economics, political science, and social science [2, 3, 31, 43]. Drawing inspiration from this foundation, we propose, TransMode-LLM, an innovative framework designed to predict travel modes from natural language descriptions of travelers and their trips. Our methodology begins with a literature-based selection of impacting factors, followed by feature importance analysis to identify the most significant predictors. We then transform these structured variables into narrative descriptions to enable LLMs to process and reason with contextual information that might be missed in traditional modeling approaches. Specifically, we introduce domain-enhanced prompting strategies and learning paradigms (zero-shot and few-shot learning approach) to enhance LLMs' understanding capabilities for travel mode prediction. To validate the proposed LLM-based approach, we conduct experiments on the real-world dateset, in which we compare our model with the state-of-the-art models with high-performance. The results demonstrate that our proposed model has the potential to yield competitive performance against established baseline models. By combining traditional travel behaviour prediction methods with the contextual intelligence of LLMs, our proposed methodology offers a novel pathway to advance travel mode prediction beyond conventional limitations.

2 RELATED WORK

2.1 Travel mode prediction

Recent years have witnessed a growing academic interest in travel mode prediction. A number of techniques have been developed on mode prediction, such as traditional statistical models and advanced machine learning models. The majority of the studies in early years on travel mode prediction focused on discrete choice models such as multinomial logit and nested logit models. These models are based on random utility theory, which estimates the probability of a traveler choosing a particular mode based on the perceived utility among alternative modes. Utility is typically constructed as a function of various attributes, such as travel time, travel cost, and comfort [4]. These models provide a theoretical foundation and interpretability, however, the principle limitations of these methods is the assumptions on the independence of irrelevant alternatives and linear-in-parameters utility functions [40], which may not adequately capture the complexity of real-world travel behavior.

To address the limitations of traditional discrete choice models, researchers have increasingly explored the application of advanced machine learning techniques in travel mode prediction. Evidence from a number of studies has confirmed the effectiveness of machine learning techniques in travel mode prediction [13, 20, 25, 36].

However, these traditional approaches including statistical models and advanced machine learning-based approaches are limited by their reliance on raw numbers and structured feature representations that may fail to capture the rich contextual nuances, complex interdependencies, and qualitative factors that may pose an impact on travel behaviour in the real-world. LLMs, can potentially address these gaps using their ability to process and reason with natural language descriptions of travel information.

2.2 Large language models

Large language models (LLMs), refer to transform-based language models containing hindered of billions or more of parameters trained on massive text data [39], such as GPT-3 [6], GPT-4 [1] and LLAMA [34]. These models demonstrate strong natural language understanding capabilities and can solve complex tasks through text generation. In recent years, researchers have shown an increased interest in applied the advanced techniques from LLMs to understand and predict human behavior across various domains including economics, political science, and social science [2, 3, 43]. Despite the successful application in various research fields, only a limited number of using LLMs on travel mode prediction have been identified [23, 26].

While prior research has demonstrated the potential of LLMs in travel behavior prediction, several significant gaps remain in the current literature. First, previous work lacks comprehensive analysis of which attributes are most influential for LLM-based travel mode prediction and how to optimally transform structured travel data into natural language descriptions that may effectively capture the nuances of travel decision-making. Second, current approaches have been evaluated on limited datasets with restricted transportation mode options that may not adequately represent the diverse travel patterns commonly used in the world, thereby raising questions about the generalizability and applicability to larger and more diverse datasets. Third, while few-shot learning has been explored, there remains insufficient investigation into how different prompting strategies i.e. incorporating enhanced domain knowledge in specialized areas—and feature representations can be beneficial to improve prediction accuracy. Finally, existing studies lack extensive comparative experiments and rigorous analysis with traditional machine learning approaches across different model configurations, dataset sizes, and evaluation metrics. This will limit the understanding of the cases that LLMs provide superior performance over traditional methods or fall behind them.

To address these gaps, we propose, TransMode-LLM, an innovative framework designed to predict travel modes from natural language descriptions of travelers and their trips. Starting from a literature-based selection of impacting factors, followed by feature importance analysis, we aim to identify the most significant attributes. We continue to propose a framework that transforms these impacting attributes into narrative descriptions to enable LLMs to process and reason with contextual information. In details, we introduce domain-enhanced prompting strategies and learning paradigms (zero-shot and few-shot learning approach) to enhance LLMs' understanding capabilities for travel mode prediction. To validate the proposed LLM-based approach, we conduct experiments on the real-world dateset, in which we compare our model with the state-of-the-art high-performance models with an establishment of comprehensive evaluation metrics.

3 METHODS

In this study, we propose a novel methodological framework for travel mode prediction that combines traditional statistical approaches and advanced natural language processing techniques. By transforming structured transportation data into rich contextual narratives processable by large language models (LLMs), we develop a comprehensive framework for travel behavior analysis with enhanced interpretability and predictive capability.

This section presents our methodology consisting of three main steps. We begin with a literature-based selection of impacting factors, followed by feature importance analysis to identify the most significant predictors. We then transform these structured variables into narrative descriptions to enable LLMs to process and reason with contextual information that might be missed in traditional modeling approaches.

3.1 Data preparation

The dataset used in this study is the National Household Travel Survey conducted by the Federal Highway Administration (FHWA), which is a periodic national survey data supporting transportation planners and policymakers in their decision-making [12]. The performance of travel behavior analysis relies on the quality of input data, particularly when developing generative and predictive models for travel mode choice. As such, we first implement a comprehensive two-stage filtering methodology to address inconsistencies within the National Household Travel Survey (NHTS) dataset that could otherwise compromise model validity.

The first step was applied a speed-distance-time consistency filter to identify physically implausible trip records such as walking trips with calculated speeds exceeding 5 mph or vehicle trips with unrealistically low speeds. Subsequently, a socio-demographic

consistency filter targeted logical inconsistencies between the characteristics of the individuals and the reported travel behaviors, such as underage drivers, non-drivers operating vehicles, young children traveling unaccompanied on public transit, and private vehicle trips without an identified driver. After filtering, the final dataset contains 22,868 trips with 85 attributes. Statistical validation demonstrates that the filtered dataset exhibits more realistic distributions across all travel modes, with particularly significant improvements in walking, cycling, and public transit categories. By addressing both physical and logical inconsistencies, these preprocessing steps significantly enhances data reliability for travel behavior analysis, enabling more accurate prediction of travel mode choice.

3.2 Preliminary analysis

Before entering the heart of the methodology by using natural language transformations to enable contextual reasoning by LLMs, we perform preliminary analysis to identify and prioritize impacting variables from the processed dataset introduced in the last section. We first identify potentially impacting factors from the literature, following by applying feature importance analysis to select the top 15 impacting factors. This two-stage selection process can help reduce the dimensionality to prevent overfitting and focuses computational resources, aiming to capture the most relevant factors of mode choice behavior.

Impacting factors of travel behavior

To perform travel mode choice prediction, a comprehensive understanding of the determinants influencing travel behavior is important. Zhou (2012) established a six-category framework for factors impacting mode choice: physical environment and urban form, mode-specific attributes, trip-makers' personal characteristics, trip parameters, travel demand management (TDM) measures, and psychological factors [42]. Building upon this foundation and drawing from the methodological approach presented by Hörl and Balac (2021) [17], this study explores five distinct categories of influential factors: individual socio-demographic characteristics, household characteristics, trip characteristics, built environment, and pricing factors. This refined categorization enables a more structured analysis of the complex interactions between personal, household, trip-specific, environmental, and economic variables that collectively shape travel behavior patterns.

- **Individual socio-demographic characteristics:**
  Individual socio-demographic characteristics play an important role in travel behaviour in many perspectives, including mode use, distances traveled, travel frequency and etc. [15]. According to the definition given by Lu et al. (1999) [24], individual socio-demographics characteristics includes age, gender, car license status and employment status. These characteristics pose a direct impact on travel behaviour and also impact travel behaviour indirectly via their impact on activity participation such as subsistence, maintenance, recreation and other. In this case, the relevant characteristics to be considered includes age, gender, driving license status and employment status.
- **Household characteristics:**
  Household information is also a very important factor on travel behaviour. A latest literature review conducted by Hu et al. (2023) indicates that there is a relationship between household information and individual activity and travel behaviors [18]. Other research also point out the similar findings [15, 24, 33]. The characteristics of household described in these studies include the number of children, number of workers, number of vehicles, household income, household type and location. These characteristics will be involved for analysis in this study.
- **Trip characteristics:**
  To comprehensively capture the characteristics of the trips, three categories are included: geospatial context, temporal patterns and engaged activities during the trips. Geospatial context refers to the origin and destination of the trip, stops during the trips as well as the travel distances. Temporal patterns include the type of the travel day (weekday or weekend), departure and arrival times, stop time and travel duration. Regarding to the engaged activities, their relevant attributes typically include the overall trip purpose, the number of stops made during the journey and corresponding purposes.
- **Built environment:**
  Built environment thought to be influencing travel mode choice have been explored in several studies [7, 10, 11, 21]. Ewing and Cervero (2010), categorized the characteristics of the built environment that impacts travel mode choice as density, diversity, design, destination accessibility, distance to transit [11]. In this analysis, key built environment variables considered include urban/rural designation (URBRUR), population size category (MSASIZE), and heavy rail availability (RAIL), which collectively capture the fundamental built environment characteristics and transportation infrastructure elements that shape travel opportunities and constraints across different geographical contexts.
- **Pricing factors:**
  Economic considerations are an important contributory factor to the travel mode choice through direct financial incentives and constraints that shape travel mode decisions. The pricing factors embedded in transportation systems, particularly fuel prices and parking costs function as market signals that directly affect the relative affordability of different travel modes. Parking costs, considered as powerful travel demand management tools that can substantially alter mode choice decisions in urban environments where alternatives to private vehicle usage exist. Several studies have explored that the parking costs have an influence on travel mode choice [16, 32, 35, 38]. Similarly, research have indicated the impact of fuel price on travel mode choice [9, 14, 19]. In this study, two key pricing variables are considered: regional gasoline prices (GASPRICE), measured in cents during the household's travel week, and parking payment status (PARK), which indicates whether travelers encountered monetary costs for vehicle storage.

To summarize, the selected variables from the preliminary analysis are summarized in Table 1.

Feature importance analysis

Table 1: Summary of factors impacting travel behavior

Category	Factor	Variable Name	Definition
Individual socio-demographic characteristics	Age	R_AGE	Respondent age
	Gender	R_SEX	Respondent sex
	Driving license status	DRIVER	Driver status
	Employment status	WORKER	Employment status of respondent
Household characteristics	Household size	HHSIZE	Total number of people in household
	Number of vehicles	HHVEHCNT	Total number of vehicles in household
	Household income	HHFAMINC	Household income
	Number of drivers	DRVRCNT	Number of drivers in the household
	Number of workers	WRKCOUNT	Count of workers in household
	Homeownership	HOMEOWN	Whether home owned or rented
Trip characteristics	Trip purpose	TRIPPURP	General purpose of trip
	Trip purpose (specific)	WHYTRP1S	Trip purpose summary
	Travel duration	TRVLCMIN	Trip Duration in Minutes
	Travel distance	TRPMILES	Calculated Trip distance converted into miles
	Day type	TDWKND	Weekend trip
	Start time	STRTTIME	24 hour local start time of trip
	End time	ENDTIME	24 hour local end time of trip
Built environment	Urban/Rural designation	URBRUR	Household in urban/rural area
	Population size	MSASIZE	Population size category of the MSA from the five-year ACS API
	Life cycle classification	LIF_CYC	Life Cycle classification for the household
	Rail availability	RAIL	MSA heavy rail status for household
Pricing factors	Gasoline price	GASPRICE	Weekly regional gasoline price, in cents, during the week of the household's travel day
	Parking costs	PARK2_PA	Amount paid for parking

To identify the most impacting factors on travel mode choice from the literature, we continue to conduct a systematic feature importance analysis. Following Cherepanova et al. (2023) which compares a number of feature-selection methods on tabular datasets [8], we apply multiple feature selection methods to overcome algorithm-specific biases and ensure robust identification of impacting factors. The analysis involves six basic methods (Univariate Statistical Test, Lasso, Random Forest, XGBoost, First-Layer Lasso, and Deep Lasso) and three advanced techniques (Adaptive Group Lasso, LassoNet, and FT-Transformer with Attention Map Importance).

The performance evaluation of our feature selection methods revealed that XGBoost achieved the highest classification accuracy (87.55%), followed by Univariate Statistical Test (86.46%) and First-Layer Lasso (86.01%). By aggregating rankings across all algorithms, we identified the top 15 most impacting factors of travel mode choice. The results demonstrate that trip distance emerges as the most consistently important impacting feature, followed by travel time, age, and gender. Other significant impacting factors include driver status, urban/rural classification, household size, household income, vehicle ownership and trip purpose.

3.3 Natural Language Description Generation

Transforming structured data into natural language descriptions is key step taken to adapt travel mode prediction for large language model processing. This transformation enables the LLM to leverage its pre-trained knowledge and contextual understanding capabilities.

The problem can be formulated as follows:

Let $X = \{x_1, x_2, ..., x_n\}$ denote the set of features for a particular trip record, where each feature x_i belongs to one of the five categories identified in our preliminary analysis: individual socio-demographic characteristics, household characteristics, trip characteristics, built environment factors, or pricing factors. We define a transformation function D that maps this feature vector to a natural language description:

$$D : X \rightarrow T \tag{1}$$

where T is the space of all possible natural language descriptions. The function D is implemented as a template-based generator that constructs coherent, contextually rich descriptions following a consistent narrative structure.

Based on our feature importance analysis, we prioritize the most influential features identified by our ensemble of feature selection methods. For each trip record, a natural language description is generated, which captures the relevant characteristics of both the

traveler and the trip. These descriptions are constructed using a consistent template that incorporated key features such as:

- Age and gender of the traveler (R_AGE, R_SEX)
- Driver and employment status (DRIVER, WORKER)
- Household characteristics (HHFAMINC, HHSIZE, HHVEHCNT, HOMEOWN)
- Trip distance, duration and purpose (TRPMILES, TRVLCMIN, TRIPPURP)
- Built environment (URBRUR, RAIL, MSASIZE)
- Pricing factors (GASPRICE)

The resulting descriptions provide a comprehensive narrative that captures the essential context of each trip while preserving the information content of the original feature vector. This approach allows the LLM to process transportation data in a format that aligns with its pre-training, enabling more nuanced interpretation of the relationships between features.

For example:

> *Consider a 44-year-old female who is a driver and is employed. She is living in a household with 3 people, and 1 vehicle, with a household income of $125,000 to $149,999, in a home that is owned with a mortgage. She is traveling for shopping, for a distance of 1.3 miles, with an expected travel time of 10 minutes. She lives in an urban area, with no access to rail transit, in an MSA of 500,000 to 999,999, where the gas price is $4.30 per gallon. What is the most likely transportation mode she would choose for this trip?*

3.4 LLM-Based Travel Mode Prediction

Problem formulation

Travel mode choice prediction can be formulated as a classification problem where the objective is to predict the mode of transportation $m \in M$ that a traveler will choose for a specific trip, given a set of features X describing personal, household, trip-specific, environmental, and economic variables that collectively shape travel behavior patterns. Therefore, this problem is represented as follows:

$$P(m|X) = f(X, \theta) \tag{2}$$

where f is a classification function with parameters θ that maps input features to probability distributions over possible transportation modes.

We propose reformulating this problem by transforming the structured feature vector X into a natural language description D, and then leveraging large language models to predict the travel mode:

$$P(m|X) \approx P(m|D(X)) \tag{3}$$

where the probability of selecting travel mode m given features X is approximated by the probability assigned by the LLM to mode m given the natural language description $D(X)$. D is a transformation function that converts structured data into a natural language description.

Prompting Strategy

The LLM-based travel mode prediction, where LLMs act as transportation analysts/planners to predict the most likely travel mode for an individual based on the contextual descriptions, represents an expansion of LLMs functionality from general tasks to domain-specific applications. The effectiveness of LLM-based travel mode prediction depends significantly on how the task is presented to the model. Using natural language to get desired responses from LLMs, known as prompting, is indeed a critical and challenging design technique in particular for domain specialization. One reason behind this is that domain-specific topics are often under-represented and involve complex concepts, terminology and relationships, making them harder to complete domain-specific tasks effectively [22]. To address this issue, we propose a *domain-enhanced prompting* to improve the effectiveness of LLMs for travel mode prediction tasks.

In detail, we incorporate standardized mode definitions inspired by the National Household Travel Survey in*domain-enhanced prompting* [5]. By conducting pre-testing via Chatbot, we find that LLMs without domain knowledge on travel mode tend to interpret 'car' as a more general vehicle concept that encompasses various passenger vehicles, while demonstrating less precise differentiation between specialized categories such as vans, SUVs/crossovers, and pickup trucks. Therefore, this strategy provides the model with domain-specific knowledge that may enhance its contextual understanding of transportation terminology.

Domain-Enhanced Prompting Example

System: You are a transportation analyst. Based on the description of a person and her/his trip, predict which transportation mode they are most likely to use from these options: Car, Van, SUV/Crossover, Pickup truck, School bus and Walk.
Car: A privately owned and/or operated licensed motorized vehicle including cars and station wagons.
SUV/Crossover: A privately owned and/or operated vehicle that is a hybrid of design elements from a van, a pickup truck, and a station wagon (e.g., Chevrolet Blazer, Ford Bronco, Jeep Cherokee).
Van: A privately owned and/or operated van or minivan designed to carry 5 to 13 passengers or to haul cargo. Pickup truck: A privately owned and/or operated motorized vehicle with an enclosed cab and an open cargo area in the rear. Typically seats 2-6 passengers.
School bus: Any bus owned, leased, or operated by a school or school district, used to transport students to/from school or related activities.
Walk: This category includes walking and jogging.
Respond with ONLY the transportation mode name (Car, SUV/Crossover, Van, Pickup truck, School bus, or Walk).
User: [Natural language description of traveler and trip]

The overview of the research framework is shown in Figure 1.

Model Selection and Comparison

The task designed for LLMs in this study focuses on predicting the most likely travel mode based on natural language descriptions which includes socio-demographic and trip-related attributes obtained from real-world survey data. To evaluate the predictive performance of LLMs for travel mode prediction, multiple LLMs varying in size and design architectures are selected to perform a systematic comparison. This comparison is motivated by the need to understand cost-effectiveness trade-offs in the adaption of LLMs

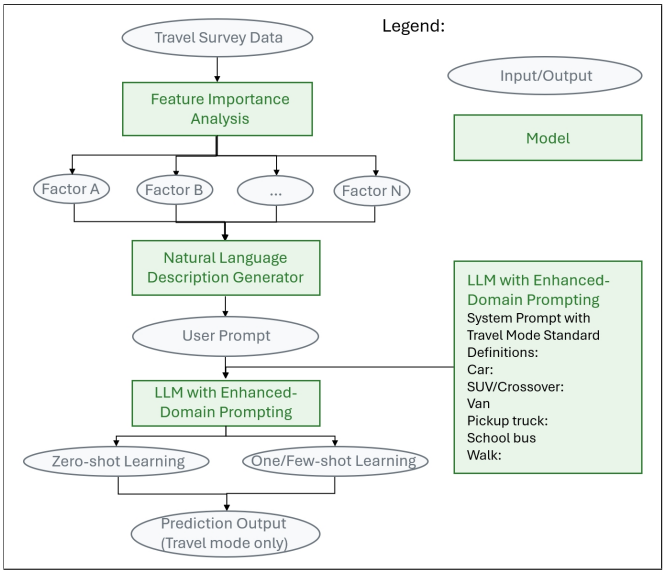

Figure 1: Research Framework of LLM-Based Travel Mode Prediction.

in travel mode prediction and transportation planning applications and also to explore whether specialized reasoning capabilities in certain models might better capture the complex decision-making processes in travel behavior.

Our selection included the full-scale GPT-4o [27] and its variant GPT-4o-mini which is designed for resource-efficient inference with reduced computational costs [28], as well as o3-mini and o4-mini, which are compact O-series reasoning models representing third and fourth generation architectures, respectively [29, 30].

With this comparative comparison among different models, it helps us to assess whether model size (e.g., GPT-4o vs GPT-4o mini) correlates directly with performance of travel mode prediction tasks, or if specialized architectures (e.g., o3 mini vs o4 mini) might offer advantages irrespective of their parameter count. In addition, this approach provides insights into the cost-effectiveness trade-offs between comprehensive and compact models for applications.

Learning Paradigms

In addition to selecting and comparing the performance of different large language models on travel mode prediction, we investigate different learning paradigms to understand their impact on predicting travel mode using natural language. Specifically, following the work established by Brown et al. [6], three learning approaches, zero-shot, one shot and few-shot learning, are considered in this study, which is categorized by the number of demonstrations provided at inference time. Specifically, zero-shot learning means prompt an LLM with no demonstration examples while one-shot and few-shot learning involve providing one or multiple demonstration examples, respectively.

Zero-shot learning in this study tests the inherent capabilities of LLMs for travel mode prediction without task-specific training. This approach leverages the pre-trained knowledge embedded within LLMs to make predictions based solely on the contextual understanding of natural language descriptions. In contrast, one-shot and few-shot learning involves providing a single and a small number of task-specific demonstrations at inference time as conditioning without permitting weight updates, respectively [6]. This approach may be advantageous for travel mode prediction, where labeled data is often scarce. In addition to one-shot learning, four more distinct few-shot configurations, providing 2, 3, 5, and 10 examples for the model to learn from, are selected to explore their impacts on the predictive performance on travel mode.

4 EXPERIMENTS

4.1 Experimental Setup

We evaluate the performance of large language models (LLMs) on the task of travel mode prediction by considering different models learning paradigms and prompting strategies. LLMs, acting as transportation analysts in this study, predict the most likely travel mode based on structured textual descriptions of individual travelers and their trip attributes. These descriptions are generated using features identified by importance analysis of the National Household Travel Survey (NHTS) dataset. In addition, to ensure practical relevance and computational efficiency, we focus on six (out of 20) travel modes: Car, Van, SUV/Crossover, Pickup truck, School bus, and Walk in this study which account for over 90% of recorded trips and coverage typical travel patterns. The experiments are implemented using Python and the OpenAI API.

4.2 Baselines

To evaluate our proposed LLM-based approach, we select high-performing statistical classifiers based on the extensive comparative analysis conducted by [37]. Their study evaluated 86 classifiers from 14 model families on the National Household Travel Survey (NHTS) 2017 dataset—structurally analogous to our dataset and identified LogitBoost and Gradient Boosting consistently superior performers across various sample sizes and datasets. We compare our proposed LLM-based approach with these two predictive models.

4.3 Implementation Details

We evaluate model performance across multiple data scales using subsets of 100, 200, 500, and 1,000 samples to assess scalability and robustness. For each sample size, we implement stratified train-test splits with an 8:2 ratio in both LogitBoost and Gradient Boosing methods. Moreover, to guarantee fair comparison between these two methods and LLM-based approaches, we use identical test sets across all methodologies within each sample size configuration. Regarding the details of LLMs' implementation, the temperature is set to 0 for models that support it (GPT-4o, GPT-4o-mini) to obtain deterministic outputs. Turning on to the details of few-shot learning approach, we examine few-shot learning effectiveness using 1, 2, 3, 5, and 10 demonstration examples. The selection of few-shot examples follows a stratified sampling strategy designed to maximize representational diversity and mode coverage.

4.4 Evaluation Metrics

We use the following commonly used metrics to evaluate the predictive performance of LLM-based approach compared to the high-performing statistical approaches mentioned in the last section.

- Accuracy: Overall percentage of correctly predicted modes
- F1-macro: Unweighted average of F1 scores for each class, measuring balanced performance across imbalanced labels.
- F1-weighted: F1 scores weighted by class support (the number of true instances for each class).

Performance assessment employs multiple metrics. Let $\hat{y}_i$ denote the predicted mode for sample i:

$$\text{Accuracy} = \frac{1}{|\mathcal{D}^{\text{test}}|} \sum_{i \in \mathcal{D}^{\text{test}}} \mathbf{1}[\hat{y}_i = y_i] \tag{4}$$

$$\text{F1}_{\text{macro}} = \frac{1}{|m|} \sum_{m \in M} \text{F1}_m \tag{5}$$

$$\text{F1}_{\text{weighted}} = \sum_{m \in M} \frac{|M_m|}{|\mathcal{D}^{\text{test}}|} \text{F1}_m \tag{6}$$

where M is the set of travel modes and F1_m is the F1-score for mode m.

Accuracy

The primary evaluation metric in this study is accuracy, the overall proportion of correct predictions compared to the actual travel mode from the dataset. Table 2 summarizes the comparative results for zero-shot learning approaches with and without domain-enhanced prompting strategies. The results demonstrate that LLMs in zero-shot learning setting show promising performance compared to traditional methods like Gradient Boosting and LogitBoost across varying sample sizes. At the smallest sample size, LogitBoost achieves exceptional performance (65.00% accuracy at n=100), significantly outperforming all LLMs. However, this advantage diminishes with the sample size increasing. The proposed LLM-based approach on travel mode prediction outperforms baseline models for larger samples. Regarding to the comparison with and without domain knowledge given, it is interesting to find that GPT-4o-mini which has a substantially smaller model size compared to the full GPT-4o model, achieves comparable accuracy to the full-scale model at the smallest sample size. It suggests that cost-efficient models can perform competitively with domain-enhanced strategy.

Interestingly, we also find that the performance advantage of domain-enhanced models over their non-domain counterparts becomes more pronounced as sample size increases, suggesting that domain knowledge becomes more valuable with more data. In addition, GPT-4o performs best at smaller sample sizes among zero-shot models, while o4-mini dominates at larger scales, indicating different models may be optimal for different tasks.

In summary, the results suggests that while baseline models have the advantage of leveraging limited data effectively, LLMs require substantial evaluation samples to demonstrate their predictive capabilities.

Turning on to few-shot learning, we focus on the best-performing few-shot learning model for each sample size and compare its accuracy with the corresponding zero-shot accuracy to evaluate the improvement. The results are summarized in Table 3. From the table, it demonstrates few-shot learning approach has the potential to improve prediction accuracy performance over zero-shot learning for LLMs. Typically requiring 2-5 examples, few-shot learning can help improve accuracy. For example, with a sample size of 100, the accuracy of the o4-mini model increases from 25.00% to 50.00%. Similar trends are observed with the o3-mini model, where accuracies improve when provided with 3 and 5 examples for sample sizes of 200 and 500, respectively. At the same time, we also find that few-shot learning does not universally outperform zero-shot approaches as expected. With a sample size of 1,000, the o4-mini model with 2 examples achieves an accuracy of 51.50% which is slightly lower than its zero-shot performance of 55.5%. This performance align with the findings in other literature [6, 41]. Zhao et al. (2020) [41] pointed out that the performance of few-shot learning seemingly minor methodological decisions such as prompt formatting, example selection, and ordering can pose a significant impact on performance, with accuracy ranging from near-chance levels to state-of-the-art performance. This area can be further investigated in the future research.

We take o3-mini as an example to provide detailed analysis and compare the performance for zero-shot learning and few-shot learning, as shown in Table 4.

The results demonstrate that few-shot learning consistently improves o3-mini's performance across all sample sizes. At smaller sample sizes (n=100), few-shot learning improves prediction accuracy from 30.0% to 45.0% (+15.0 percentage points) with just 2 given examples. Moreover, the improvement from few-shot learning exhibits an inverse relationship with sample size. The optimal number of examples varies across sample sizes, with no clear monotonic relationship. Small sample sizes (n=100, n=200) benefit from multiple examples (2-5), while larger samples achieve optimal performance with fewer examples (1-3). This pattern suggests that as the evaluation set grows larger, the model may already have sufficient context to perform well, requiring minimal additional guidance from few-shot examples.

In summary, these findings reveal a complex trade-off between model complexity, data requirements, and performance consistency, with important implications for practical deployment in transportation planning applications where data availability and prediction accuracy requirements vary significantly.

F1 Scores

Table 5 presents the comparative results of F1-macro and F1-weighted scores varying by LLM architectures (GPT-4o, GPT-4o-mini, o3-mini, and o4-mini) and learning paradigms (zero-shot and few-shot) across different sample sizes. From the table, it is obvious that o4-mini demonstrates superior F1-weighted performance (0.5101) of zero-shot learning approach with a sample size of 1,000, underscoring its effectiveness in prioritizing majority travel modes, while its F1-macro score, 0.4826 in the same sample size under 3-shot learning approach indicates robust handling of minority travel modes, suggesting it a versatile model for balanced performance.

In contrast, gpt-4o performs stable but less divergent F1-macro and F1-weighted scores, peaking at 0.4481 and 0.4099 respectively under 5-shot learning approach with a sample size 1,000. This suggests gpt-4o model consistent but less specialized behavior across travel mode distributions. Furthermore, the significant improvement in gpt-4o-mini's F1-macro (from 0.2143 to 0.3985 in the case of 1,000 sample size) under few-shot learning highlights the efficacy of example-based learning for enhancing minority travel mode prediction, though its modest F1-weighted gains indicate limited impact on majority travel modes. It is also worthy noting that in

Table 2: Accuracy of Models under Zero-shot learning across Sample Sizes

Sample Size	Gradient Boosting	LogitBoost	Best Zero-shot Learning Models			
			Without Domain		Domain-enhanced	
			Accuracy	Best Model	Accuracy	Best Model
100	0.4500	**0.6500**	0.4000	GPT-4o	0.4000	GPT-4o-mini
200	0.4750	0.4750	0.5000	GPT-4o	**0.5500**	o4-mini
500	0.5300	0.4500	**0.5400**	o4-mini	0.5300	o4-mini
1,000	**0.4800**	**0.4800**	0.5250	o4-mini	**0.5550**	o4-mini

Table 3: Accuracy of Models under Few-shot Learning across Sample Sizes

Sample Size	Best Few-shot Learning Models			
	Few-shot Accuracy	Best Model (Examples)	Zero-shot Accuracy	Improvement
100	0.5000	o4-mini (3)	0.2500	+0.2500 (+100.0%)
200	0.5000	o3-mini (5)	0.3750	+0.1250 (+33.3%)
500	0.5000	o3-mini (3)	0.4100	+0.0900 (+18.0%)
1,000	0.5150	o4-mini (2)	0.5550	-0.0400 (-7.2%)

Table 4: Zero-shot vs Few-shot Performance for o3-mini across Sample Sizes

Sample Size	Zero-shot Accuracy	Few-shot Accuracy	No. of Examples	Improvement (%)
100	0.3000	0.4500	2	+15.0
200	0.3750	0.5000	5	+12.5
500	0.4100	0.5000	3	+9.0
1,000	0.4350	0.4750	1	+4.0

the case with larger sample sizes, the convergence of F1-macro and F1-weighted scores across most models reflects improved class balance handling, yet o4-mini's persistent divergence suggests a unique architectural bias toward majority travel mode optimization. In summary, these findings highlight the importance of aligning model selection and shot configuration with specific evaluation metrics, thereby providing valuable guidance and suggestions for applying LLMs in knowledge discovery tasks in the future where class imbalance is a critical factor.

5 CONCLUSION

This study proposes a novel framework using the contextual reasoning capabilities of large language models (LLMs) to predict travel mode choice from natural language descriptions of travel survey data. In this study, we transform structured features into richly descriptive textual inputs including sociodemographic, trip, and built environment and economic contexts, bridging the gap between traditional quantitative modeling and emerging language-based AI techniques. Furthermore, we introduce domain-enhanced prompting strategies and learning paradigms (zero-shot and few-shot learning) to enhance LLMs' understanding capabilities for travel mode prediction. By conducting extensive experiments across varying sample sizes using the real-world dataset,the results demonstrate that this LLM-based approach has the potential to yield competitive performance against established baseline models, even in low-data regimes through the application of proposed domain-enhanced prompting strategy and few-shot learning. In addition, to evaluate the performance of the proposed LLM-based approach comprehensively, three commonly used metrics (Accuracy, F1-macro score and F1-weighted) are selected in this study. These findings highlight the dual potential of large language models in transportation research: not only as powerful predictive tools but also as interpretable frameworks that can provide insights into the multifaceted nature of travel decisions, which also offers significant advantages for transportation planning practitioners and policymakers who require explainable model outputs for decision-making processes.

This research opens several promising avenues for future investigation. Advanced prompt engineering techniques, such as chain-of-thought reasoning, could further improve model performance and interpretability. Moreover, domain-specific pretraining of language models on transportation field could yield specialized models better attuned to the nuances of travel behavior modeling.

6 ACKNOWLEDGMENT

This research was partially funded by a grant from the Singapore Ministry of Education (MOE-T2EP20122-0012).

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In

Table 5: F1-macro and F1-weighted scores across different LLMs with varying numbers of examples and sample sizes

Model	No. of Examples	F1-Macro				F1-Weighted			
		100	200	500	1000	100	200	500	1000
gpt-4o	0	0.2020	**0.2354**	0.1935	0.1982	0.2303	**0.3096**	0.2620	0.2789
	1	**0.3909**	0.2352	0.3310	0.3318	**0.2936**	0.3750	0.2995	0.3370
	2	0.3611	**0.3734**	0.3095	0.3907	0.2667	**0.4023**	0.2917	0.3612
	3	0.3655	0.2590	0.3357	**0.4030**	0.2746	0.3717	0.3347	**0.3779**
	5	0.3840	0.2570	0.3056	**0.4481**	0.3245	0.3516	0.4157	**0.4099**
	10	0.2063	0.2214	0.2313	**0.4069**	0.2794	0.3679	0.3696	**0.3629**
gpt-4o-mini	0	**0.2500**	0.2390	0.2125	0.2143	**0.2500**	0.3021	0.2785	0.2892
	1	0.2203	0.2111	0.2329	**0.3306**	0.2365	0.2917	0.2896	**0.3147**
	2	0.2203	0.1951	0.2044	**0.3985**	0.2365	0.3096	0.2717	**0.3192**
	3	**0.3909**	0.1548	0.1732	0.3217	**0.2936**	0.2304	0.2503	0.3057
	5	**0.3092**	0.1696	0.1947	0.2518	**0.2565**	0.2714	0.2912	0.2788
	10	**0.3730**	0.1512	0.3112	0.3373	**0.2881**	0.2355	**0.3205**	0.3011
o3-mini	0	0.0800	0.0926	0.1733	**0.2555**	0.1440	0.2083	0.2596	**0.2865**
	1	0.2020	0.2132	0.2192	**0.3426**	0.2303	0.2963	0.2844	**0.3201**
	2	**0.2613**	0.1714	0.2508	0.3463	**0.3325**	0.3000	0.3346	0.3236
	3	0.1968	0.2132	0.2657	**0.3536**	0.1943	0.2963	0.3862	**0.3206**
	5	0.2333	0.2643	0.3399	**0.4155**	0.2800	**0.3979**	0.3211	**0.3702**
	10	0.3526	0.2413	0.3527	**0.4345**	0.2247	0.3525	0.3365	**0.3945**
o4-mini	0	0.2340	0.2744	0.3409	**0.4177**	0.2394	0.3653	**0.4824**	**0.5101**
	1	**0.2702**	0.2126	0.2580	0.3609	**0.3663**	0.2949	0.3494	0.3842
	2	0.2281	0.2605	0.2700	**0.4631**	0.2839	0.3688	0.3741	**0.4383**
	3	0.3693	0.2681	0.2981	**0.4826**	**0.4794**	0.3845	0.3873	**0.4325**
	5	**0.3725**	0.2413	0.2666	0.3965	**0.2806**	0.3525	0.3403	0.4323
	10	**0.2641**	0.2413	0.2632	0.3928	**0.3518**	0.3525	0.3514	0.4206

International Conference on Machine Learning. PMLR, 337–371.
[3] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis* 31, 3 (2023), 337–351.
[4] Moshe E Ben-Akiva and Steven R Lerman. 1985. *Discrete choice analysis: theory and application to travel demand*. Vol. 9. MIT press.
[5] Stacey Bricka, Timothy Reuscher, Paul Schroeder, Mitchell Fisher, Justina Beard, Xiaoyuan Layla Sun, et al. 2024. *Summary of Travel Trends: 2022 National Household Travel Survey*. Technical Report. United States. Federal Highway Administration. Office of Policy and ....
[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
[7] Cynthia Chen, Hongmian Gong, and Robert Paaswell. 2008. Role of the built environment on mode choice decisions: additional evidence on the impact of density. *Transportation* 35 (2008), 285–299.
[8] Valeriia Cherepanova, Roman Levin, Gowthami Somepalli, Jonas Geiping, C Bayan Bruss, Andrew G Wilson, Tom Goldstein, and Micah Goldblum. 2023. A performance-driven benchmark for feature selection in tabular deep learning. *Advances in Neural Information Processing Systems* 36 (2023), 41956–41979.
[9] Felix Creutzig. 2014. How fuel prices determine public transport infrastructure, modal shares and urban form. *Urban climate* 10 (2014), 63–76.
[10] Chuan Ding, Donggen Wang, Chao Liu, Yi Zhang, and Jiawen Yang. 2017. Exploring the influence of built environment on travel mode choice considering the mediating effects of car ownership and travel distance. *Transportation Research Part A: Policy and Practice* 100 (2017), 65–80.
[11] Reid Ewing and Robert Cervero. 2010. Travel and the built environment: A meta-analysis. *Journal of the American planning association* 76, 3 (2010), 265–294.
[12] Federal Highway Administration. 2022. 2022 NextGen National Household Travel Survey Core Data. http://nhts.ornl.gov. Accessed: 2024-12-09.
[13] Julian Hagenauer and Marco Helbich. 2017. A comparative study of machine learning classifiers for modeling travel mode choice. *Expert Systems with Applications* 78 (2017), 273–282.
[14] Ashley R Haire and Randy B Machemehl. 2007. Impact of rising fuel prices on US transit ridership. *Transportation Research Record* 1992, 1 (2007), 11–19.
[15] Susan Hanson and Perry Hanson. 1981. The travel-activity patterns of urban residents: dimensions and relationships to sociodemographic characteristics. *Economic geography* 57, 4 (1981), 332–347.
[16] Daniel Baldwin Hess. 2001. Effect of free parking on commuter mode choice: Evidence from travel diary data. *Transportation Research Record* 1753, 1 (2001), 35–42.
[17] Sebastian Hörl and Milos Balac. 2021. Synthetic population and travel demand for Paris and Île-de-France based on open and publicly available data. *Transportation Research Part C: Emerging Technologies* 130 (2021), 103291.
[18] Yang Hu, Bert van Wee, and Dick Ettema. 2023. Intra-household decisions and the impact of the built environment on activity-travel behavior: A review of the literature. *Journal of Transport Geography* 106 (2023), 103485.
[19] Boris Jäggi, Alexander Erath, Christoph Dobler, and Kay W Axhausen. 2012. Modeling household fleet choice as function of fuel price by using a multiple discrete–continuous choice model. *Transportation Research Record* 2302, 1 (2012), 174–183.
[20] Mohammad Tamim Kashifi, Arshad Jamal, Mohammad Samim Kashefi, Meshal Almoshaogeh, and Syed Masiur Rahman. 2022. Predicting the travel mode choice with interpretable machine learning techniques: A comparative study. *Travel Behaviour and Society* 29 (2022), 279–296.
[21] Joonho Ko, Sugie Lee, and Miree Byun. 2019. Exploring factors associated with commute mode choice: An application of city-level general social survey data. *Transport policy* 75 (2019), 36–46.
[22] Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, et al. 2023. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *arXiv preprint arXiv:2305.18703* (2023).

[23] Tianming Liu, Manzi Li, and Yafeng Yin. 2024. Can large language models capture human travel behavior? evidence and insights on mode choice. *Evidence and Insights on Mode Choice (August 26, 2024)* (2024).
[24] Xuedong Lu and Eric I Pas. 1999. Socio-demographics, activity participation and travel behavior. *Transportation Research part A: policy and practice* 33, 1 (1999), 1–18.
[25] Yixuan Ma and Zhenji Zhang. 2020. Travel mode choice prediction using deep neural networks with entity embeddings. *IEEE Access* 8 (2020), 64959–64970.
[26] Baichuan Mo, Hanyong Xu, Dingyi Zhuang, Ruoyun Ma, Xiaotong Guo, and Jinhua Zhao. 2023. Large language models for travel behavior prediction. *arXiv preprint arXiv:2312.00819* (2023).
[27] OpenAI. 2024. GPT-4o. https://platform.openai.com/docs/models/gpt-4o. Accessed: 2025-04-22.
[28] OpenAI. 2024. GPT-4o mini. https://platform.openai.com/docs/models/gpt-4o-mini. Accessed: 2025-04-22.
[29] OpenAI. 2024. o3-mini. https://platform.openai.com/docs/models/o3-mini. Accessed: 2025-04-22.
[30] OpenAI. 2024. o4-mini. https://platform.openai.com/docs/models/o4-mini. Accessed: 2025-04-22.
[31] Shohreh Shaghaghian, Luna Yue Feng, Borna Jafarpour, and Nicolai Pogrebnyakov. 2020. Customizing contextualized language models for legal document reviews. In *2020 ieee international conference on big data (big data)*. IEEE, 2139–2148.
[32] Donald Shoup. 2021. *High cost of free parking*. Routledge, New York.
[33] Sumeeta Srinivasan and Joseph Ferreira. 2002. Travel behavior at the household level: understanding linkages with residential choice. *Transportation Research Part D: Transport and Environment* 7, 3 (2002), 225–242.
[34] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
[35] Natasa Vidovic and Jelena Simicevic. 2023. The impact of parking pricing on mode choice. *Transp. Res. Procedia* 69, 3 (2023), 297–304.
[36] Fangru Wang and Catherine L Ross. 2018. Machine learning travel mode choices: Comparing the performance of an extreme gradient boosting model with a multinomial logit model. *Transportation Research Record* 2672, 47 (2018), 35–45.
[37] S Wang, B Mo, and J Zhao. 2020. Predicting travel mode choice with 86 machine learning classifiers: An empirical benchmark study. In *Proc. 99th Annu. Meeting Transp. Res. Board*. 279–296.
[38] Kevin Washbrook, Wolfgang Haider, and Mark Jaccard. 2006. Estimating commuter mode choice: A discrete choice analysis of the impact of road pricing and parking charges. *Transportation* 33 (2006), 621–639.
[39] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* 1, 2 (2023).
[40] Xilei Zhao, Xiang Yan, Alan Yu, and Pascal Van Hentenryck. 2020. Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. *Travel behaviour and society* 20 (2020), 22–35.
[41] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*. PMLR, 12697–12706.
[42] Jiangping Zhou. 2012. Sustainable commute in a car-dominant city: Factors affecting alternative mode choices among university students. *Transportation research part A: policy and practice* 46, 7 (2012), 1013–1029.
[43] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics* 50, 1 (2024), 237–291.