Exploring the Potential of Large Language Models in Daily Travel Activity Pattern Prediction

Mahan Mollajafari*[1], Zachary Patterson[2], and Bilal Farooq[3]

[1]PhD Student, Information and Systems Engineering, Concordia University, Canada
[2]Professor, Concordia Institute for Information Systems Engineering, Canada
[3]Associate Professor, Laboratory of Innovations in Transportation, Toronto Metropolitan University, Canada

Short summary

Daily Travel Activity Patterns (DAPs/DTAPs) are critical in transportation modelling, especially within Activity-Based frameworks, enabling planners and policymakers optimize network performance and operational efficiency. While multi-label classification approaches have been previously applied for DAP prediction, the comparison of Large Language Models (LLMs) with other widely used algorithms in this domain remains unexplored. Leveraging the strengths of LLMs in sequence modelling and multi-label classification, this paper makes a novel contribution by implementing and comparing several open-access LLMs for DAP prediction against traditional discrete choice models (Multinomial Logit) and popular machine learning and deep learning algorithms. Using the 2018 Origin-Destination Montreal travel survey data, including socio-demographic information on over 169,000 individuals, results demonstrate that LLMs can improve prediction accuracy by up to 3% over other methods. However, this improvement comes with significantly higher training times, highlighting a trade-off between accuracy and computational efficiency.
Keywords: Daily Travel Activity Pattern, Activity Based Models, Large Language Models, Artificial Intelligence, Discrete Choice Models

1 Introduction

Daily travel activity patterns (DAPs/DTAPs), which describe the sequence and timing of daily activities, are central to transportation studies, particularly in Activity-Based Modelling (ABM) frameworks (Bowman and Ben-Akiva, 2001). Analysing DAPs reveals factors shaping travel behaviour, essential for predicting travel demand and developing effective transportation policies (Kitamura, 1988). Traditional trip-based models, such as the widely used four-stage model (trip generation, distribution, mode choice, and assignment), predict transportation demand by analysing trip origins and destinations (Bhat and Koppelman, 1999). These models assume independent travel decisions, ignoring activity interdependencies, time constraints, and family dynamics (Bhat and Koppelman, 1999; Pas, 1985), limiting their ability to capture complex travel behaviour and leading to suboptimal policies (Bowman and Ben-Akiva, 2001).

To address these limitations, ABMs focus on individual activity patterns and decision-making processes, considering socioeconomic factors, land use, and transportation attributes (Bhat and Koppelman, 1999; Gärling et al., 1994). This enables more realistic representations of human mobility, improving transportation policies and interventions (Arentze and Timmermans, 2004; Bowman and Ben-Akiva, 2001). Some notable ABMs are Bowman's Boston ABM (Bowman and Ben-Akiva, 1997), CEMDAP (Bhat et al., 2004), ALBATROSS (Arentze and Timmermans, 2000), TASHA (Roorda et al., 2008), CUSTOM (Nurul Habib, 2018), and SALT (Hafezi et al., 2021). Generally, ABMs are classified into rule-based, utility-based, and hybrid models. Rule-based models use heuristic rules and behavioural theories like bounded rationality to simulate travel decisions under constraints such as time availability and activity preferences (Arentze and Timmermans, 2000; Recker, 1995). Utility-based models, grounded in utility maximization, employ discrete choice methods (DCMs) like the Multinomial Logit (MNL) and Nested Logit (NL) to predict activity-travel behaviour (Ben-Akiva and Lerman, 1985; Allahviranloo and Recker, 2013). Hybrid models combine rule-based and utility-based approaches, offering more comprehensive and realistic simulations and improved predictions for transportation planning and policy analysis (Arentze and Timmermans, 2004; Bhat and Koppelman, 1999).

The rise of Machine Learning (ML) and Deep Learning (DL) techniques has significantly advanced DAP modelling by enabling the capturing of complex, non-linear patterns from large datasets. Unlike previous models, ML and DL approaches adapt to changing behaviour and offer more flexible and accurate predictions (Gong et al., 2014). Therefore, researchers have been increasingly drawn to use these techniques with diverse data sources, such as GPS, smartphones, travel surveys, and smart card data. While these models excel at handling complex data, their "black-box" nature and lack of interpretability remain significant challenges, especially when compared to the transparency offered by traditional DCMs. ML algorithms like the Random Forest (RF), Support Vector Machines (SVM), and Gradient Boosting (GB) have demonstrated strong predictive capabilities for DAPs. Hafezi et al. (2018) used RF models to capture activity dependencies and socio-demographic heterogeneity, achieving high accuracy. Nayak and Pandit (2023) applied RF, XGBoost, and LightGBM within a multi-label framework, modelling interdependencies between weekday and weekend activities with improved validation results. Allahviranloo and Recker (2013) employed SVMs and Hidden Markov Models to capture sequential dependencies, outperforming traditional MNL models. Deng (2022) modelled senior DAPs using boosted C5.0 algorithms, offering both accuracy and interpretability through surrogate rule-based models. In addition, DL techniques, such as neural networks and transformer-based models, further enhance DAP prediction. Wang and Osaragi (2024) showed that the Time-Varying Markov Chain (TVMC) achieved accuracy comparable to Multi-Layer Neural Networks (MNN) while offering greater interpretability. Phan and Vu (2021) used DL frameworks with entity embedding and domain knowledge to classify activities and predict activity times, demonstrating the potential of DL for reliable DAP generation.

Natural Language Processing (NLP) techniques are also excel at sequence and pattern recognition, particularly in multi-label classification tasks (Tsoumakas and Katakis, 2007). They effectively handle ordered data, such as activity sequences and time series, by uncovering contextual relationships and extracting features from raw data. Scalable and adaptable, NLP methods are well-suited for analysing large, complex transportation datasets (Raaijmakers, 2022; Vaswani et al., 2017). Li and Lee (2017) developed probabilistic context-free grammars (PCFGs) to generate DAPs, effectively capturing the complexity of activity sequences. Chen et al. (2024) combined NLP-based feature extraction (Word2Vec+SIF) with clustering techniques (K-Means++) and ML algorithms to identify detailed activity patterns. Their approach demonstrated consistency across years and robustness for long-term predictions. Artificial Intelligence (AI) models, including ML, DL, and NLP, outperform DCMs in prediction tasks by handling large datasets, capturing non-linear patterns, and achieving higher accuracy, particularly for DAP recognition (Wang et al., 2021). Although, these models face a "black box" problem, limiting their interpretability, this can often be overlooked since prediction accuracy and training efficiency often draw more attention.

While significant progress has been made in using AI algorithms for DAP modelling and prediction, the implementation and performance comparison of various open-access Large Language Models (LLMs) in domain has not yet been investigated, despite their success in sequence modelling and pattern recognition. It should be noted that traditional sequence models, such as Recurrent Neural Networks (RNNs) (Graves et al., 2013) and Long Short-Term Memory (LSTM) networks Hochreiter and Schmidhuber (1997), excel at capturing temporal dependencies but require large sequential datasets for effective training and often face challenges like vanishing gradients, making them unsuitable for this task (Hochreiter and Schmidhuber, 1997). In contrast, LLMs, offer significant advantages over traditional NLP techniques. LLMs are pre-trained on vast corpora using advanced transformer architectures, enabling them to model long-range dependencies, capture rich contextual relationships, and perform well with smaller datasets through fine-tuning (Chen et al., 2024; Vaswani et al., 2017; Devlin et al., 2018). Their ability to capture complex dependencies in data, whether through bidirectional contextual embeddings or unidirectional sequence modelling, makes them particularly effective for tasks like multi-label classification (Niraula et al., 2024).

The purpose of this study is to implement and compare a range of open-access LLMs, including both bidirectional and unidirectional models, using multi-label classification approaches to predict individual DAPs based on socio-demographic information. The 2018 Origin-Destination (OD) Montreal Travel Survey dataset, which contains personal, household, and travel information for approximately 169,000 individuals, is used as a case study. To evaluate the performance of LLMs, a comprehensive comparison is conducted against ML algorithms, including RF, SVM, and GB, as well as a DL model (MNN) and a DCM (MNL). Additionally, the relationships between different activities, previously analysed using techniques like Hidden Markov Models (Allahviranloo and Recker, 2013) or Markov chains (Wang and Osaragi, 2024), are modelled using classifier chains, which are well-suited for multi-label classification problems.

2 Methodology

The purpose of this study is to explore the ability to improve DAP prediction by using LLM models compared to existing methods. The framework's steps are described in detail below.

Case Study

The 2018 OD Montreal Travel Survey was chosen for its comprehensive data on daily travel behaviours (Autorité régionale de transport métropolitain, 2018). Conducted from September to December 2018, it covered households in Greater Montreal to support planning and infrastructure development. Stratified sampling divided the area into 113 strata based on 2016 census data, ensuring demographic and spatial representation. Figure 1 shows the dispersion of surveyed household locations. The survey gathered data from 73,400 households, capturing 357,798 valid movements from 168,883 individuals. Its large sample size offers a significant advantage over previous studies. This analysis focuses on individuals making 0 to 5 daily trips, representing 95% of the sample (163,122 individuals). Figure 2 displays the histogram of individuals with various daily trip numbers. Finally, dataset variables are grouped into Household, Personal, and Movement categories. For model training, commonly used variables relevant to DAP prediction were selected. Table 1 details the chosen variables, including their categories, names, descriptions, and types.

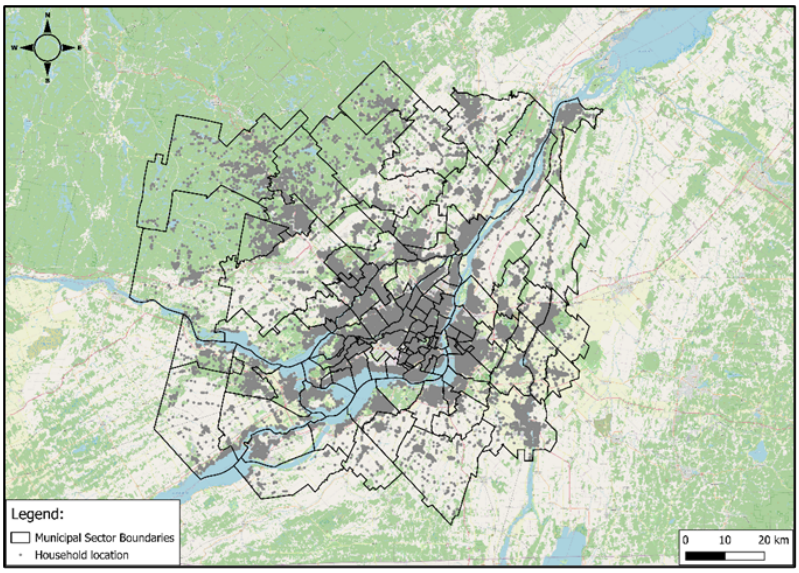

Figure 1: Dispersion of the recorded household locations in the dataset

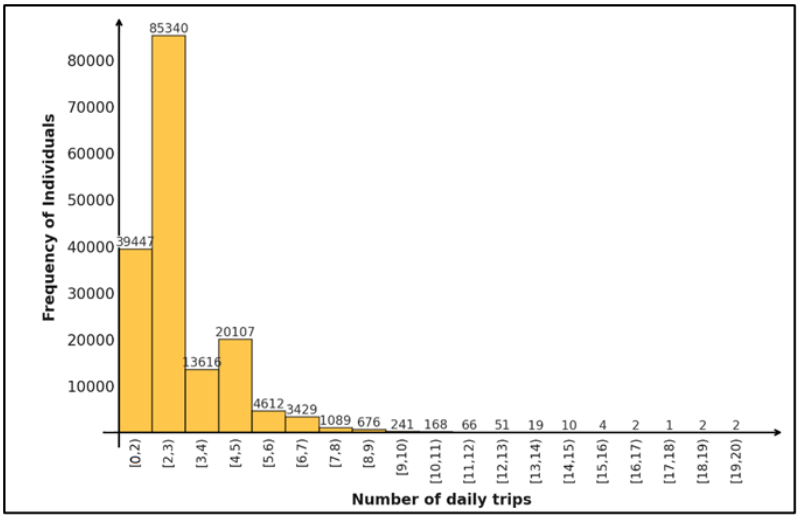

Figure 2: Histogram of the recorded daily trip numbers of individuals in the dataset

Table 1: Information of the Chosen variables in the dataset

Category	Feature	Feature Description	Type
Household	m_fexp	Expansion factor based on households	Continues Numerical
	m_auto	Number of vehicles in the household (0 to 14)	Discrete Numerical
	m_pers	Number of people in the household (1 to 19)	Discrete Numerical
	m_domsm	Municipal sector (113 municipal sectors)	Categorical
	m_domlon	Longitude of the household location	Continues Numerical
	m_domlat	Latitude of the household location	Continues Numerical
Personal	p_fexp	Weighting factor based on people	Continues Numerical
	p_sexe	Sex of the person (1: male, 2: female)	Categorical
	p_grage	Age group of the person (1: 0 to 4 years, 2: 5 to 9 years, 3: 10 to 14 years, 4: 15 to 19 years, 5: 20 to 24 years, 6: 25 to 34 years, 7: 35 to 44 years, 8: 45 to 54 years, 9: 55 to 64 years, 10: 65 to 74 years, 11: 75 and older)	Categorical
	p_age	Age of the person (1 to 99)	Discrete Numerical
	p_statut	Main occupation of the person (1: Full-time worker, 2: Part-time worker, 3: Student/pupil, 4: Retired, 5: Other, 6: N/A: child under 4 years old, 7: At home, 8: Refusal)	Categorical
	p_permis	Possession of a driver's license: (1: Yes, 2: No, 3: Don't know, 4: Refusal, 5: Not applicable (under 16 years old))	Categorical
	p_mobil	Mobility of the person: (1: Yes, 2: No, did not move, 3: N/A: child under 4 years old, 4: Don't know, 5: Refusal, 6: Moved, don't know how)	Categorical
Movement	num_trps	Number of daily trips	Discrete Numerical
	d_motif	Reason for the movement: (Category 0- No trip/staying home [0: No trip], Category 1- Mandatory activities [1: Work, 4: School], Category 2- Maintenance activities [2: Business meeting, 5: Shopping, 8: Health], Category 3- Discretionary activities [3: On the road, 6: Leisure, 7: Visit friends/relatives, 9: Look for someone, 10: Pickup someone, 12: Other], Category 4- Home [11: Return to Home])	Categorical

DAP preparation and representation

Since DAPs are not directly recorded in the dataset, they are represented as target variables for modelling using two approaches: pur_pat_0 and pur_pat_1. The pur_pat_0 representation includes 13 detailed trip purpose categories, enabling granular analysis, while pur_pat_1 simplifies these into 5 generalized groups: (0) no trip, (1) mandatory, (2) maintenance, (3) discretionary, and (4) return home (Bhat and Koppelman, 1999; Phan and Vu, 2021), as shown in Table 1. The simplified pur_pat_1 approach supports behavioural insights and travel demand modelling by standardizing trip purposes and highlighting patterns such as the rigidity of work trips versus the flexibility of leisure trips (Allahviranloo and Recker, 2013; Hafezi et al., 2018; Chen et al., 2024). In this research, only individuals with fewer than six daily trips are included. Therefore, DAPs are represented as fixed-size vectors of 5 labels, with values depending on the chosen representation. For instance, an individual with four trips categorized as shopping, visiting friends, leisure, and returning home would have a DAP vector of [5, 7, 6, 11, 0] in pur_pat_0 and [2, 3, 3, 4, 0] in pur_pat_1. The reduced complexity and fewer categories in pur_pat_1 are expected to result in higher prediction accuracies compared to pur_pat_0.

Model Selection

The purpose of this study is to evaluate a range of open-access LLMs, including both bidirectional models such as BERT (Devlin et al., 2018), DistilBERT (Sanh et al., 2019), ALBERT (Lan et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), and T5 (Raffel et al., 2020), as well as unidirectional models such as BLOOM (Le Scao et al., 2023), Falcon (Gao et al., 2024), and LLaMA 2 (Touvron et al., 2023).
These transformer-based models process sequential data through tokenization, embedding layers, and attention mechanisms to capture local and global relationships. Bidirectional models, such as BERT, use tasks like Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) for pre-training (Devlin et al., 2018), capturing both past and future context. Unidirectional models, like LLaMA 2, employ causal (auto-regressive) learning to predict the next token, offering faster and more efficient computation (Touvron et al., 2023). During fine-tuning, task-specific layers adapt these models for multi-label classification problems, enabling effective processing of sequential data for diverse applications. Each LLM has unique strengths, reflecting a trade-off between contextual understanding and computational efficiency. Bidirectional models excel in interpreting complex activity relationships, while unidirectional models are more efficient for tasks with sequential dependencies or computational constraints.
Finally, widely used ML models like RF, SVM, and GB are included for their proven suitability in DAP prediction, along with the DL model MNN to capture non-linear relationships. The benchmark MNL model is also used for its interpretability and decision-making focus. This evaluation contextualizes LLM performance against existing models, highlighting their potential advantages.

Preprocessing and Training the models

Preprocessing varies by algorithm. For ML models and MNN, categorical and numerical features are standardized using *OneHotEncoder* and *StandardScaler* from *scikit-learn* (Scikit-learn developers, 2023). For LLMs, text features are tokenized using *transformers* library tokenizers (HuggingFace, 2023) and converted into numerical format. For the MNL model, alternatives and input variables are defined using *Biogeme* library (Bierlaire, 2023).
The dataset is split into train (70%), validation (15%), and test (15%) sets, with validation used to monitor over-fitting and test for final evaluation. Sequential trip relationships are modelled using classifier chains, where previous predictions inform subsequent ones, enhancing accuracy despite increased training time.
Training employs *scikit-learn* for ML models, *TensorFlow* (Abadi et al., 2016) for MNN, *transformers* for LLMs, and *Biogeme* for MNL. The hyper-parameters are optimized using *GridSearchCV* for ML models and MNN, and *Optuna* (Team, 2023) for LLMs. MNL variables are refined based on t-test significance and coefficients, excluding low-significance or non-meaningful coefficients variables for better reliability and interpretability.

Model Evaluation

This study proposes two evaluation approaches: vector-based accuracy and unit-based accuracy. Since the target is a vector of categorical labels, categorical cross-entropy is used for accuracy calculations. In the vector-based approach, all labels in an individual's vector must be predicted correctly; a single incorrect label results in the entire vector being considered a false prediction. In contrast, unit-based accuracy calculates the mean correctness of predicted labels within a DAP. For example, if two out of five labels are incorrect, the vector's accuracy would be 0% in the vector-based approach but 60% in the unit-based approach.
These evaluation methods assess model performance based on the order of trips, trip purpose categories, and individuals with varying numbers of daily trips, under the two mentioned DAP representation methods. Since unit-based accuracy is more lenient, it is expected to yield higher values than vector-based accuracy. A summary of the framework is provided in Figure 3.

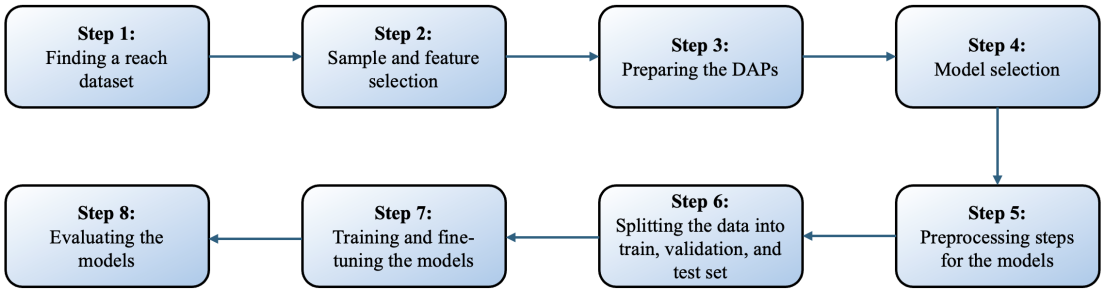

Figure 3: Proposed framework used in this study

3 Results and Discussion

This ongoing project requires additional time to implement T5, BLOOM, Falcon, and LLaMA 2 due to their high training costs and numerous scenarios, which will be analysed in the coming months. Consequently, the current analysis focuses on the remaining algorithms. Table 2 compares their performance, assessing runtime and accuracy (vec_acc and unit_acc) under two DAP representation methods. The evaluation includes total observations, trip order, trip purpose categories, and daily trip counts, with pur_pat_1 highlighted in some scenarios due to word limits. Notably, MNL was not implemented for pur_pat_0 due to its computational complexity, underscoring its limitations with more complex representations.

Running time: Training times vary notably between pur_pat_1 and pur_pat_0, reflecting differing complexities. SVM is the fastest overall, completing pur_pat_1 in 32.88 seconds and pur_pat_0 in 81.24 seconds, while MNL is the slowest, requiring 42,096.96 seconds and even longer for pur_pat_0. Among LLMs, DistilBERT is the quickest (8,608.8 seconds for pur_pat_1 and 12,169.58 seconds for pur_pat_0), with XLNet the slowest.

Total Accuracy: The results highlight three key trends: (1) all models achieve higher accuracy under pur_pat_1 than pur_pat_0, reflecting the reduced complexity of the simpler representation; (2) unit-based accuracy (unit_acc) surpasses vector-based accuracy (vec_acc), as vec_acc requires perfect label prediction, making it more challenging; and (3) LLMs, particularly RoBERTa, deliver higher accuracy in most scenarios due to their advanced contextual modelling.
For vec_acc (Figure 4), RoBERTa consistently leads, achieving 74.10% for pur_pat_1 and 68.35% for pur_pat_0. For unit_acc (Figure 5), models perform better overall, with RoBERTa scoring highest under both representations (92.77% for pur_pat_1 and 91.15% for pur_pat_0). While RoBERTa and other LLMs achieve marginally better total accuracy, traditional models like MNN and GB offer slightly lower performance but significantly faster training. This highlights a trade-off between training time and accuracy, making traditional AI models more efficient for certain applications.

Based on the order of the trips: Figure 6 illustrates model accuracy (unit_acc) by trip order under pur_pat_1, with accuracy rising from 79–82% for the first trip to over 99% by the fifth. RoBERTa starts and finishes with the highest accuracy, exceeding 99%. MNN outperforms others for the second, third, and fourth trips, achieving 91.67%, 93%, and 98.73%, respectively, while RoBERTa remains competitive. Accuracy improves across all models as trip order increases, driven by narrowing trip category distributions and classifier chains leveraging logical relationships between trips.

Table 2: Performance of the models in different scenarios

DAP Representation method	Evaluation Metric		Number of observations	Model									
				LLM					ML			DL	Discrete Choice
				BERT	ALBERT	DistilBERT	RoBERTa	XLNet	RF	SVM	GB	MNN	MNL
pur_pat_1	Run_Time		24469 (all)	19871.28	11612.16	8608.8	19577.28	26003.64	106.92	**32.88**	286.44	433.44	42096.96
	Vec_Acc		24469	73.40%	73.75%	73.60%	**74.10%**	73.50%	71.47%	70.53%	72.53%	72.47%	71.45%
	Unit_Acc		24469	92.64%	92.66%	92.61%	**92.77%**	92.52%	91.77%	90.55%	92.59%	92.51%	92.51%
	Based on the order of the trips	Trip 1	24469	81.40%	81.30%	81.50%	**81.85%**	81.25%	80.07%	79.93%	79.60%	80.47%	79.25%
		Trip 2	24469	90.85%	91.00%	90.90%	91.00%	90.50%	90.47%	88.20%	**91.67%**	90.20%	91.50%
		Trip 3	24469	92.60%	92.65%	92.30%	92.65%	92.50%	90.20%	85.73%	**93.00%**	93.00%	92.85%
		Trip 4	24469	98.40%	98.40%	98.40%	98.40%	98.40%	98.20%	98.93%	98.73%	**98.93%**	99.05%
		Trip 5	24469	99.95%	99.95%	99.95%	**99.95%**	99.95%	99.93%	99.93%	99.93%	99.93%	99.90%
	Based on the trip categories	Cat 0	74489	100.00%	100.00%	100.00%	**100.00%**	100.00%	99.98%	100.00%	100.00%	100.00%	100.00%
		Cat 1	13058	87.45%	88.29%	87.92%	88.38%	86.52%	90.34%	**91.12%**	89.30%	89.69%	88.34%
		Cat 2	5618	37.45%	43.10%	37.24%	42.68%	**47.70%**	33.14%	39.05%	40.53%	44.08%	46.55%
		Cat 3	7783	57.93%	53.51%	58.69%	**54.42%**	50.76%	44.88%	34.20%	46.41%	43.14%	44.38%
		Cat 4	21400	98.47%	98.17%	97.76%	**98.53%**	98.23%	95.17%	89.80%	97.93%	97.47%	97.99%
	Based on the number of daily trips (vec_acc)	0-1 trips	5898	96.48%	95.86%	96.69%	**96.69%**	95.86%	96.61%	94.52%	96.08%	96.61%	91.49%
		2 trips	12762	82.71%	82.71%	82.71%	**82.80%**	82.71%	79.23%	80.52%	79.23%	80.26%	80.06%
		3 trips	2087	34.12%	39.41%	34.71%	**39.41%**	35.29%	23.21%	0.89%	32.14%	17.86%	29.33%
		4 trips	3038	32.40%	33.20%	33.20%	33.60%	34.00%	31.31%	35.35%	34.34%	**37.37%**	34.62%
		5 trips	684	3.57%	1.79%	3.57%	3.57%	1.79%	0.00%	3.13%	**6.25%**	3.13%	4.26%
	Based on the number of daily trips (unit_acc)	0-1 trips	5898	99.30%	99.17%	99.34%	**99.34%**	99.17%	99.32%	98.90%	99.22%	99.32%	98.30%
		2 trips	12762	96.48%	96.48%	96.48%	**96.50%**	96.50%	95.77%	96.03%	95.74%	95.97%	95.92%
		3 trips	2087	81.76%	82.00%	81.76%	**82.24%**	80.59%	72.68%	53.21%	81.25%	77.14%	79.47%
		4 trips	3038	79.52%	79.84%	79.76%	79.92%	80.32%	80.20%	81.11%	80.51%	**81.82%**	81.77%
		5 trips	684	55.36%	55.00%	52.86%	**56.07%**	51.79%	43.13%	46.88%	51.25%	46.88%	56.17%
pur_pat_0	Run_Time		24469	22850.45	21623.84	12169.58	23431.98	29511.27	156.60	**81.24**	1091.30	496.06	NA
	Vec_Acc		24469	68.00%	68.15%	68.15%	**68.35%**	68.30%	66.07%	66.33%	66.67%	66.93%	
	Unit_Acc		24469	90.83%	90.94%	90.94%	**91.15%**	90.97%	90.01%	89.88%	90.49%	90.85%	
	Based on the order of the trips	Trip 1	24469	78.75%	78.95%	78.95%	**79.20%**	79.00%	76.93%	77.47%	76.13%	78.00%	
		Trip 2	24469	88.60%	88.50%	88.50%	88.80%	88.60%	88.93%	88.20%	**89.60%**	89.13%	
		Trip 3	24469	88.95%	89.05%	89.05%	**89.55%**	89.05%	86.60%	85.53%	88.47%	88.73%	
		Trip 4	24469	97.90%	98.25%	98.25%	98.25%	98.25%	97.80%	98.27%	98.33%	**98.47%**	
		Trip 5	24469	99.95%	99.95%	99.95%	**99.95%**	99.95%	99.80%	99.93%	99.93%	99.93%	
	Based on the number of daily trips (vec_acc)	0-1 trips	5898	95.24%	95.24%	95.24%	96.27%	96.27%	**96.61%**	95.56%	95.82%	95.56%	
		2 trips	12762	79.35%	79.35%	79.35%	**79.35%**	79.25%	74.97%	76.65%	74.58%	76.90%	
		3 trips	2087	16.47%	16.47%	16.47%	16.47%	15.88%	6.25%	0.00%	**16.96%**	6.25%	
		4 trips	3038	18.40%	**19.60%**	19.60%	19.20%	18.40%	16.67%	17.68%	17.68%	17.68%	
		5 trips	684	0.00%	0.00%	0.00%	0.00%	**5.36%**	0.00%	0.00%	3.13%	0.00%	
	Based on the number of daily trips (unit_acc)	0-1 trips	5898	99.05%	99.05%	99.05%	99.25%	99.25%	**99.32%**	99.11%	99.16%	99.11%	
		2 trips	12762	95.81%	95.81%	95.81%	**95.81%**	95.79%	94.92%	95.25%	94.81%	95.30%	
		3 trips	2087	76.00%	76.00%	76.00%	**76.12%**	75.88%	68.39%	63.21%	75.89%	73.93%	
		4 trips	3038	73.60%	73.68%	73.68%	74.56%	73.68%	73.74%	74.14%	73.23%	**74.75%**	
		5 trips	684	49.29%	52.86%	52.86%	**54.29%**	52.86%	36.25%	40.00%	40.00%	43.13%	

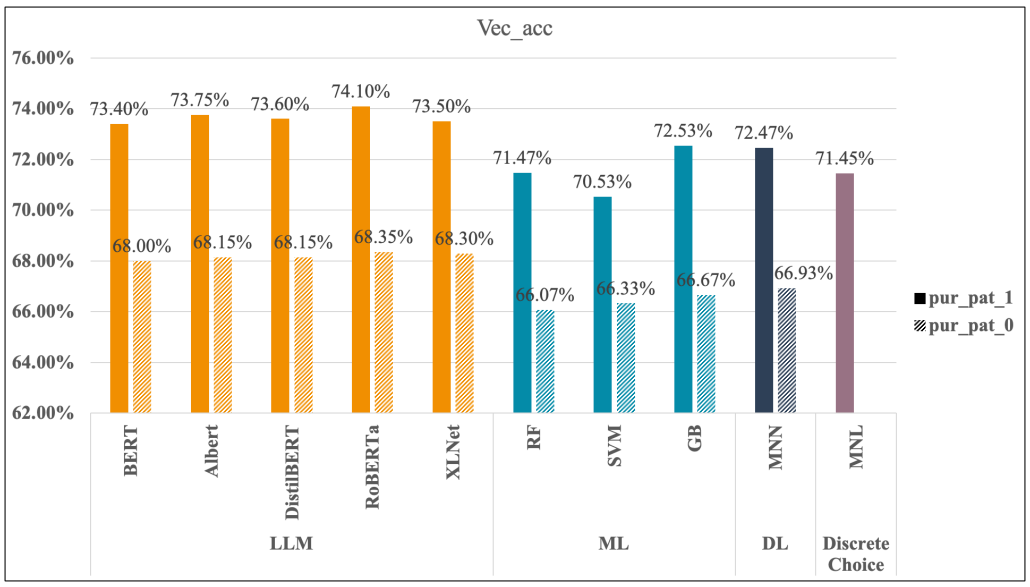

Figure 4: Total vec_acc of the models

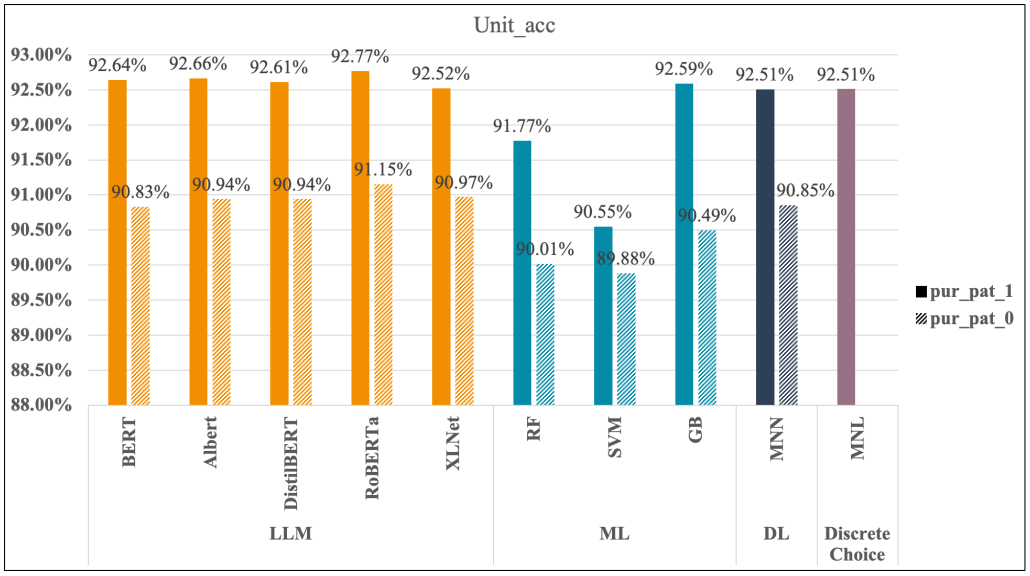

Figure 5: Total unit_acc of the models

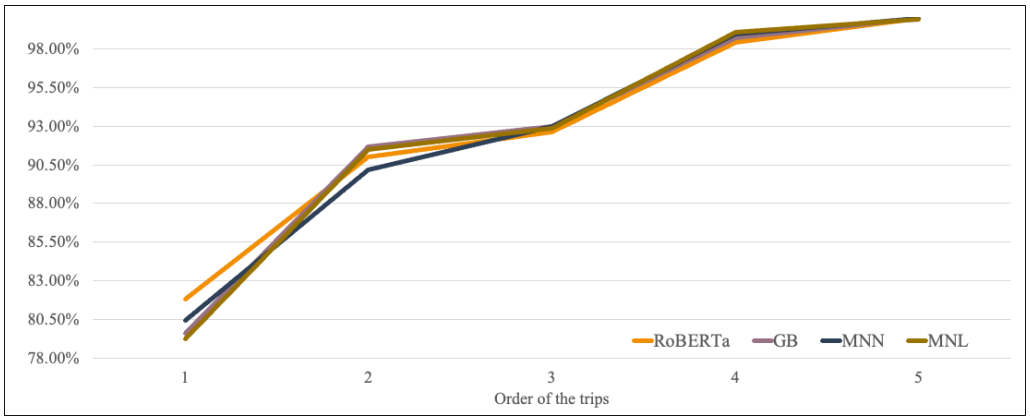

Figure 6: unit_acc of the models based on the order of the trips under pur_pat_1

Based on the trip categories: Figure 7 shows unit_acc by trip purpose under pur_pat_1. All models achieved perfect accuracy for "No Trip." MNN led in "Mandatory activities" (89.69%), MNL in "Maintenance" (46.55%), and RoBERTa in "Discretionary" (54.42%) and "Home" (98.53%). Lower observation counts for "Maintenance" and "Discretionary" activities (Table 2) posed challenges for all models. RoBERTa stood out as the best overall, excelling in "No Trip," "Discretionary," and "Home," while remaining competitive in other categories.

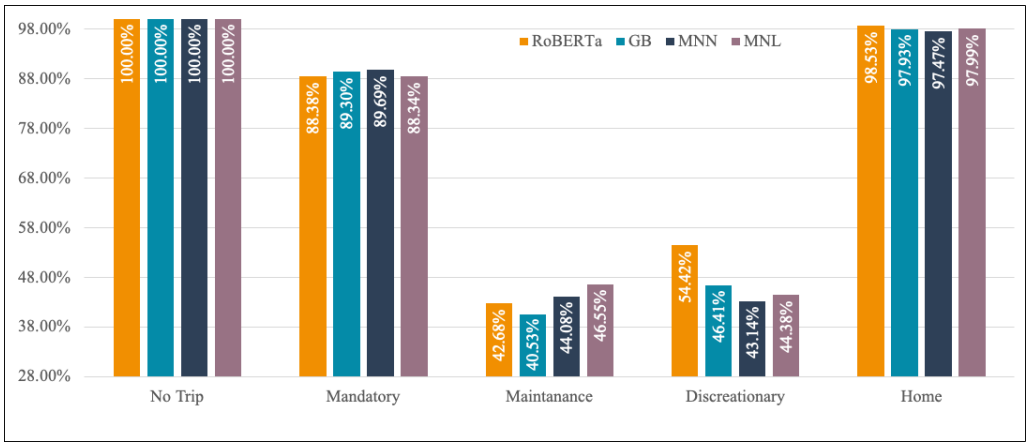

Figure 7: unit_acc of the models based on the trip purposes under pur_pat_1

Based on the daily trip number: Figures 8 and 9 display vec_acc and unit_acc by individuals' daily trip numbers (0-1, 2, 3, 4, 5 trips) under pur_pat_1. In vec_acc, all models perform well for "0-1 trips," with RoBERTa leading at 96.69%. Accuracy declines as trips increase, with RoBERTa highest for "2 trips" (82.80%) and "3 trips" (39.41%), MNN best for "4 trips" (37.37%), and GB for "5 trips" (6.25%). In unit_acc, RoBERTa excels for all trip numbers except "4 trips," achieving 99.34% for "0-1 trips," 96.50% for "2 trips," 82.24% for "3 trips," and 56.07% for "5 trips." MNN leads for "4 trips" with 81.82%. Accuracy decreases for "3 trips" and "5 trips" due to fewer observations (Table 2). Both graphs highlight RoBERTa's dominance and the general decline in accuracy as trip numbers rise.
In summary, selecting an algorithm for DAP prediction involves balancing accuracy, training time, and interpretability. RoBERTa achieves the highest accuracy but has long training times, while traditional ML and DL models offer faster training with slightly lower accuracy. AI models outperform MNL in both accuracy and speed, making them more efficient for most applications. For interpretability, however, MNL remains the only option despite its high computational cost. This highlights the trade-off between precision, efficiency, and interpretability in algorithm selection.

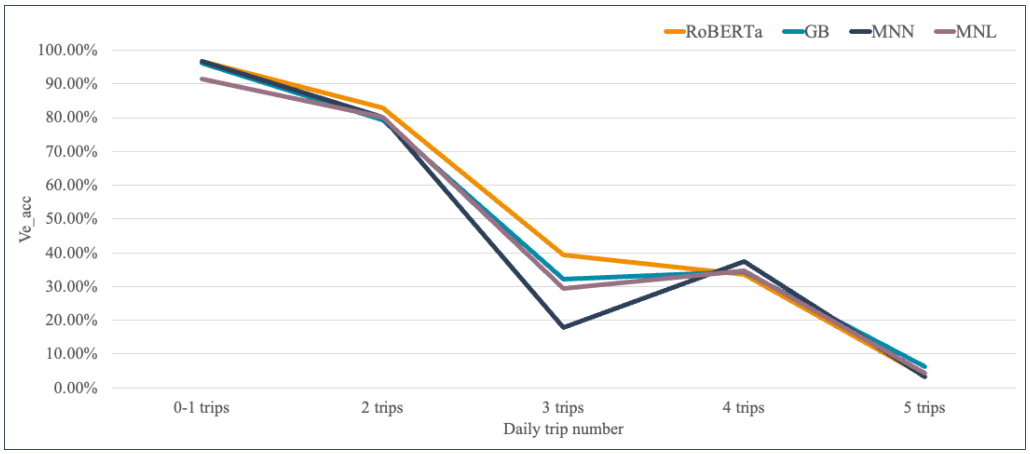

Figure 8: vec_acc of the models based on the daily trip numbers under pur_pat_1

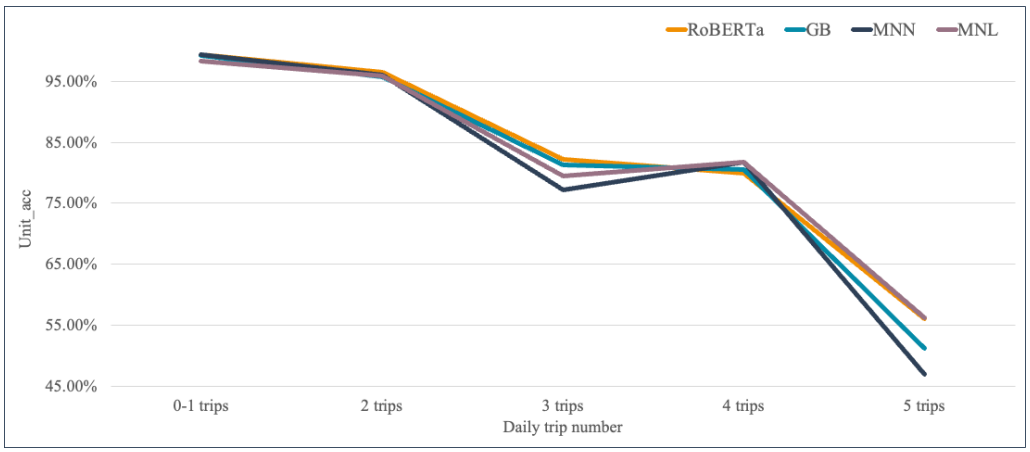

Figure 9: unit_acc of the models based on the daily trip numbers under pur_pat_1

4 Conclusions

This study explored the use of several open-source LLMs for predicting DAPs of approximately 169,000 individuals using socio-demographic data from the 2018 OD Montreal Travel Survey. The key contribution lies in a systematic and comprehensive comparison of open-source LLMs with traditional theory-driven and AI-driven ML and DL models, using a multi-label classification approach that to the best of our knowledge, has not been previously applied to LLMs. DAPs represented by two methods (pur_pat_1 and pur_pat_0), focusing primarily on pur_pat_1 for its reduced complexity. Classifier chains were used to model trip relationships, and models were assessed using vector-based and unit-based accuracies.

Among all models, RoBERTa achieved the highest predictive accuracy across most scenarios, outperforming both traditional models and other LLMs, considering that T5, BLOOM, Falcon, and LLaMA 2, will be analysed in the upcoming months. However, the improvement over traditional AI models was marginal, up to 3% under pur_pat_1 and 2% under pur_pat_0, respectively. DCMs like MNL are interpretable but require extensive training time and simplification, limiting scalability. ML and DL models are faster and accurate but lack interpretability. LLMs, while slower, offer marginally better accuracy due to advanced contextual modelling. The choice of algorithm depends on project priorities: interpretability, efficiency, or accuracy.

For future research, it is crucial to evaluate these models on diverse datasets to capture variations in activity patterns across different cities. Further studies should explore alternative approaches, such as modelling the entire DAP vector as a single target instead of relying on classifier chains. Integrating LLMs with additional datasets—such as geospatial, sensor, or social media data—or combining them with advanced methodologies like survival analysis or competing risk models could significantly enhance DAP prediction accuracy. Additionally, examining temporal dynamics, including seasonal variations and long-term trends, offers a promising direction for understanding and improving activity pattern recognition.

Acknowledgements

The authors would like to thank the Regional Transportation Authority (ARTM), Ministry of Transportation of Quebec (MTQ), Ministry of Municipal Affairs and Housing (MAMH), and other local transit bodies for a collaborative effort for gathering and preparing the 2018 OD Montreal Travel Survey. This research is funded by Canada First Research Excellence Fund (CFREF) funded Bridging Divides research program.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Allahviranloo, M. and Recker, W. (2013). Daily activity pattern recognition by using support vector machines with multiple classes. *Transportation Research Part B: Methodological*, 58:16–43.

Arentze, T. and Timmermans, H. (2000). *Albatross: a learning based transportation oriented simulation system*. Citeseer.

Arentze, T. A. and Timmermans, H. J. (2004). A learning-based transportation oriented simulation system. *Transportation Research Part B: Methodological*, 38(7):613–633.

Autorité régionale de transport métropolitain (2018). Enquête Origine-Destination 2018.

Ben-Akiva, M. E. and Lerman, S. R. (1985). *Discrete choice analysis: theory and application to travel demand*, volume 9. MIT press.

Bhat, C. R., Guo, J. Y., Srinivasan, S., and Sivakumar, A. (2004). Comprehensive econometric microsimulator for daily activity-travel patterns. *Transportation Research Record*, 1894(1):57–66.

Bhat, C. R. and Koppelman, F. S. (1999). A retrospective and prospective survey of time-use research. *Transportation*, 26:119–139.

Bierlaire, M. (2023). Biogeme: A free software for the estimation of discrete choice models. https://biogeme.epfl.ch/. Accessed: 2024-12-16.

Bowman, J. L. and Ben-Akiva, M. (1997). Activity-based travel forecasting. In *Activity-Based Travel Forecasting ConferenceDepartment of TransportationFederal Transit AdministrationFederal Highway AdministrationOffice of the Secretary of TransportationEnvironmental Protection Agency*.

Bowman, J. L. and Ben-Akiva, M. E. (2001). Activity-based disaggregate travel demand model system with activity schedules. *Transportation research part a: policy and practice*, 35(1):1–28.

Chen, M., Yuan, Q., Yang, C., and Zhang, Y. (2024). Decoding urban mobility: Application of natural language processing and machine learning to activity pattern recognition, prediction, and temporal transferability examination. *IEEE Transactions on Intelligent Transportation Systems*.

Deng, Y. (2022). Application of machine learning with a surrogate model to explore seniors' daily activity patterns. *Transportation letters*, 14(9):972–982.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gao, X., Xie, W., Xiang, Y., and Ji, F. (2024). Falcon: Faster and parallel inference of large language models through enhanced semi-autoregressive drafting and custom-designed decoding tree. *arXiv preprint arXiv:2412.12639*.

Gärling, T., Kwan, M.-p., and Golledge, R. G. (1994). Computational-process modelling of household activity scheduling. *Transportation Research Part B: Methodological*, 28(5):355–364.

Gong, L., Morikawa, T., Yamamoto, T., and Sato, H. (2014). Deriving personal trip data from gps data: A literature review on the existing methodologies. *Procedia-Social and Behavioral Sciences*, 138:557–565.

Graves, A., Mohamed, A.-r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee.

Hafezi, M. H., Daisy, N. S., Millward, H., and Liu, L. (2021). Framework for development of the scheduler for activities, locations, and travel (salt) model. *Transportmetrica A: Transport Sci*, 18(2):248–280.

Hafezi, M. H., Liu, L., and Millward, H. (2018). Learning daily activity sequences of population groups using random forest theory. *Transportation research record*, 2672(47):194–207.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

HuggingFace (2023). Transformers: State-of-the-art natural language processing for tensorflow 2.0 and pytorch. Accessed: 2023-07-28.

Kitamura, R. (1988). An evaluation of activity-based travel analysis. *Transportation*, 15:9–34.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., et al. (2023). Bloom: A 176b-parameter open-access multilingual language model.

Li, S. and Lee, D.-H. (2017). Learning daily activity patterns with probabilistic grammars. *Transportation*, 44:49–68.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nayak, S. and Pandit, D. (2023). A joint and simultaneous prediction framework of weekday and weekend daily-activity travel pattern using conditional dependency networks. *Travel Behaviour and Society*, 32:100595.

Niraula, N., Ayhan, S., Chidambaram, B., and Whyatt, D. (2024). Multi-label classification with generative large language models. In *2024 AIAA DATC/IEEE 43rd Digital Avionics Systems Conference (DASC)*, pages 1–7. IEEE.

Nurul Habib, K. (2018). A comprehensive utility-based system of activity-travel scheduling options modelling (custom) for worker's daily activity scheduling processes. *Transportmetrica A: Transport Science*, 14(4):292–315.

Pas, E. I. (1985). State of the art and research opportunities in travel demand: another perspective. *Transportation Research A*, 19(5/6):460–464.

Phan, D. T. and Vu, H. L. (2021). A novel activity pattern generation incorporating deep learning for transport demand models. *arXiv preprint arXiv:2104.02278*.

Raaijmakers, S. (2022). *Deep learning for natural language processing*. Simon and Schuster.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Recker, W. W. (1995). The household activity pattern problem: General formulation and solution. *Transportation Research Part B: Methodological*, 29(1):61–77.

Roorda, M. J., Miller, E. J., and Habib, K. M. (2008). Validation of tasha: A 24-h activity scheduling microsimulation model. *Transportation Research Part A: Policy and Practice*, 42(2):360–375.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Scikit-learn developers (2023). scikit-learn: Machine Learning in Python. Accessed: 2023-07-28.

Team, O. (2023). Optuna: A hyperparameter optimization framework. Accessed: 2023-07-28.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tsoumakas, G. and Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, S., Mo, B., Hess, S., and Zhao, J. (2021). Comparing hundreds of machine learning classifiers and discrete choice models in predicting travel behavior: an empirical benchmark. *arXiv preprint arXiv:2102.01130*.

Wang, W. and Osaragi, T. (2024). Generating and understanding human daily activity sequences using time-varying markov chain models. *Travel Behaviour and Society*, 34:100711.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.