Deciphering Travel Choices: A Comparative Analysis of Large Language Models and Traditional Decision-Making Frameworks

Jung-Hoon Cho, Youry Moise, Gigi Sung, Hanyong Xu*

HIGHLIGHTS

- The study compares travel mode choice predictions of traditional decision-making frameworks such as that of Expected Utility Theory (EUT) with the prediction capacity of Large Language Models (LLMs).
- The study demonstrates that LLMs can nearly match the predictive performance of traditional models like EUT, despite not being exposed to specific training examples.
- The study identifies three effective strategies for enhancing the predictive accuracy of LLMs in travel mode choice: incorporating diverse features, creating behavior-focused prompts, and fine-tuning the models with human reasoning data.

ABSTRACT

This study analyzed how Large Language Models (LLMs) make judgments in the context of the travel mode choice decision-making process and existing behavioral studies, and compared it with a traditional decision-making framework. Utilizing the Swissmetro dataset, we first applied Expected Utility Theory via a Multinomial Logistic Model to establish a benchmark for travel mode choice predictions. We then established three sets of experiments with GPT-3.5 and compared the results with the benchmark model. First, we analyzed the impact of feature choice on LLM performance and reasoning. Second, we investigated the value of role prompting in terms of human behaviors. Lastly, we surveyed students and requested reasonings on travel behaviors and utilized them to fine-tune the LLM to make predictions. We found that in general, enriching the model with a broader range of features, designing prompts that reflect behavioral tendencies, and fine-tuning the model using human-like reasoning significantly enhance the predictive capabilities of GPT-3.5. We achieved a result almost comparable to that of Expected Utility Theory trained on a large dataset without showing GPT any training examples. This provides crucial evidence and directs potential future research in operationalizing LLMs to make travel behavior decision-making predictions that could benefit transportation planners.

I. INTRODUCTION

Since last year, Large Language Models (LLMs) have generated tremendous excitement in research for their advanced language interaction and interpretation capabilities exhibited by products such as ChatGPT. In transportation, studies have shown the abilities of GPT to make mobility predictions [1]. However, the question of whether LLMs formulate a similar reasoning process as humans when they make decisions is unclear. In the context of mode choice decision-making using LLMs, it is crucial to explore this process before one can safely and efficiently apply it to user products.

II. PROBLEM STATEMENT

A. Research questions and hypotheses

Understanding human travel behavior has always been a crucial part of transportation planning. Travel mode choice is one of the characteristics that describe a person's travel behavior, which refers to how people make decisions on their means of travel. This research will focus on analyzing how LLM infers

* The authors contributed equally. Alphabetical order.

travel mode choice decisions for a person given relevant background information based on its reasoning outputs. Specifically, the study will answer the following research questions:

- What kind of variable inputs impact its output reasonings, and in what way? Is it comparable to the Expected Utility Theory?
- Does the way of describing the travel mode choice context matter for LLM? In other words, does prompting LLMs to assume rational or behavioral agents change the prediction outcome?
- How does human intuition play a role in helping LLM to make better decisions?

B. Significance and purposes

Since many companies are developing solutions to utilize the power of LLM in user products to make mobility recommendations to people, this project can help us understand the trustworthiness of LLMs by analyzing their reasoning limitations or potentials in the context of travel mode decision-making.

III. Literature Review

A. Large Language Models

Large Language Models (LLMs) typically refer to the language models based on the Transformer architecture with more than hundreds of billions of parameters which have been shown to have abilities to excel in many complex tasks [2]. Wei et al. [3] have demonstrated the emerging abilities of LLMs that only appear in larger models and cannot be predicted based on smaller models' performance. Newer models such as GPT-4 illustrated strong capabilities in language processing and manipulation, quantitative reasoning, as well as planning and learning [4]. Products like ChatGPT which implemented LLMs into chatbots and search engines that are easily accessible by the general public generated tremendous attention and excitement in this field [2].

While there is rarely a study that specifically addresses LLMs in the context of predicting human travel behavior, some sources that relate to the topics of LLMs predicting human behavior and displaying biases in decision-making help us understand the capability and limitation of LLMs and thus design an effective experiment.

Before proceeding, it is essential to consider how human bias can influence LLMs. Agrawal et al. (2023) point out that while LLMs are proficient in making predictions, they inherently lack the ability to exercise judgment in decision-making [5]. Such judgment requires an understanding of individual preferences and contexts, an area where LLMs fall short. To address this, "reward function engineers" play a crucial role in incorporating human judgment into LLMs. This approach is evident in models like ChatGPT, which have been fine-tuned with human feedback to yield outputs that are contextually relevant, safe, and of high quality. This integration of human judgment significantly boosts the practical utility of LLMs in various applications. However, it simultaneously introduces a vulnerability to human biases, as these models are shaped by human input and perspectives.

Meanwhile, a study by MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) demonstrates that LLMs can exhibit biases similar to human decision-making and explores logic as a means to mitigate these biases [6]. They experimented with logic-aware models trained to discern relationships between sentences based on context and semantics, effectively reducing harmful stereotypes compared to baseline models. For example, a logic-trained model neutrally interprets the relationship between "the person is a doctor" and "the person is masculine," countering the stereotype-based correlations in common models.

While these results do not directly address whether LLMs exhibit behaviors like those in Prospect Theory or their ability to infer human choices in travel mode selection, they highlight the stereotyped nature of LLMs and suggest methods to develop less biased LLMs for transportation applications.

B. Decision Making Frameworks

1) Expected Utility Theory: Commonly used as a framework for understanding how people make decisions, Expected Utility Theory (EUT) claims that, given decisions with uncertain outcomes associated with certain payoffs, agents will act rationally, acting in a way that maximizes their expected utility. EUT calculates this value using the traditional expected value formula. We take all the potential outcomes of making a decision, multiply them by their probabilities, find their sum, and see if the result is greater than the result of making a different decision. If it is, we choose that option. In other words, we want to make the choice that maximizes $\Sigma_i p_i u_i$ over each associated outcome i. An important underlying assumption of this model is that we, when faced with uncertainty, are able to accurately judge the probability of any event taking place. This raises concerns because, as we will discuss, humans tend to predictably overweight or underweight certain probabilities, causing our estimates of expected utility to be incorrect.

An important distinction from behavior-facing theories such as Prospect Theory is that this model uses a utility function that depends on both the potential change in utility and the current level of wealth. As a result, two agents faced with the same situation may be predicted to make different decisions because of different levels of wealth. It also does not consider the framing or risk-averse nature of the individual making the decision.

2) Prospect Theory: An alternative, more widely accepted human behavior theory is Prospect Theory. It builds on two important ideas about our decision-making process.

Firstly, we experience a phenomenon called the endowment effect, where, after making any kind of decision, regardless of how we felt prior to doing so, we start to become attached to the outcome of that decision. This is what drives campaigns such as money-back guarantees and free trials because businesses know that, after purchasing, people will be reluctant to return the product due to their emotional attachment to it [7].

Secondly, Prospect Theory relies on the idea that there is often not one objective way to present choices. We present using gain framing (emphasizing the benefits of either situation) or loss framing (emphasizing what we stand to lose). Tversky and Kahneman found that, under the former, people tend to be risk averse, while under the latter, people are more risk-taking [8].

We also consider that we, as humans, typically cannot calculate probabilities with 100% accuracy. Specifically, we overestimate low probabilities and underestimate high probabilities, which means we need a function to calculate the perceived probability when getting our expectations. The function in question is $\pi(P) = \frac{P^\gamma}{(P^\gamma+(1-P)^\gamma)^{\frac{1}{\gamma}}}$ where P is the probability of an outcome. We find the value of each outcome with $v = x^\alpha$ for gains, and $v = -\lambda(-x)^\alpha$ for losses. Here, x is the change in wealth, and λ is the coefficient of loss aversion, indicating the extent to which the individual wants to avoid losses. For each decision, we find $\Sigma_i v_i \pi_i$ for each outcome i, and choose the decision with the highest summation.

C. Multinomial Logistic Regression

A popular tool for discrete choice prediction models is multinomial logistic regression, which is meant for predicting the value of a categorical dependent variable based on the values of several independent variables[9]. It is essentially traditional logistic regression, but capable of handling outcome variables with more than two categories.

A significant advantage is the few assumptions required to perform this kind of regression. Unlike other models, we do not make many critical assumptions about the underlying distribution of the data. Firstly, there is no assumption of normality, meaning multinomial regression works whether the data follows a normal distribution or not. For other methods, such as t tests, break down completely if the data is, in reality, not normal. Secondly, we do not assume linearity, which means the relationship between the independent and dependent variables, if one exists, does not need to be linear. Finally, we do not need homoscedasticity, which states that variance should remain consistent for all values of an

independent variable. For example, within one variable, values of lower magnitude should not yield lower residuals than those of higher magnitude. Without this assumption, there are many models that yield skewed results, but it is not a problem for multinomial regression. However, there are some key assumptions that must be satisfied. The first is independence among the alternatives of the dependent variable. Secondly, each independent variable has one, and only one, value for each case.

Finally, we need non-perfect separation, meaning the outcome variable cannot clearly separate any of the independent variables. An example of a dataset that fails this requirement would be one in which there is some constant, c, such that, anytime the outcome variable takes on its first value, the predictor variable is less than c, and when the outcome variable takes on its other value, the predictor variable is greater than c.

There are other important constraints that must be met, such as not having outliers and all continuous variables being linear. After getting our coefficients, we must decide whether they indicate a true relationship between the predictors and the outcome variables using a hypothesis test. Our null hypothesis is that there is no relationship between the supposed predictors and the outcome variable, which would yield coefficients of 0. Our alternate hypothesis is that there is a relationship between these variables. Once we have the coefficients from our model, we calculate the probability of getting them, given that the null hypothesis was true. If this probability is less than 0.05, we reject the null hypothesis, determining that the coefficients imply a relationship between our variables. Because the probabilities involved are typically incredibly small, we use the log-likelihood (the natural logarithm of the probabilities) instead of the probabilities themselves. When running the model, we start at iteration 0, where we find the log-likelihoods associated with the null model, in which none of the predictors are included. Afterward, we introduce the predictors and, for every iteration, make adjustments to the coefficients in an attempt to decrease the log-likelihood. Once the differences in logs between iterations become relatively small, we conclude that the model has "converged" and return the coefficients associated with this model.

IV. Methodology

As discussed in the problem statement, this research aims to examine how LLM makes decisions compared to that of traditional models, such as the Expected Utility Theory, and how to improve their decision-making process based on behavioral theories. To explore this, As shown in 1, we first established the Utility Function based on the Swissmetro dataset as the benchmark model. Then, we examined and compared the efficacy of Large Language Models (LLMs), specifically ChatGPT 3.5 (referred to as 'Vanilla GPT') against the benchmark in predicting travel mode choice decisions. Moreover, we also tested how framing the prompt as "rational" or "Prospect Theory aware" can affect the predictions.

In the second part of the study, we extended our analysis through a custom survey that captures travel mode choices and reasonings under three distinct scenarios: Daily Commute, Weekend Trip, and Business Travel. The survey findings were then used to fine-tune the Vanilla GPT model. The enhanced model was subsequently tested against the Swissmetro dataset to evaluate whether the fine-tuned GPT model offers better explanatory power for people's travel mode choices. To analyze how different features might play a role in the performance, both models are trained with combinations of three different sets of variables: plain (basic travel information), demographics, and other factors (such as luggage).

A. Data

The Swissmetro dataset [10] was employed to design and compare the decision-making processes of both LLM and EUT models. The data was gathered from surveys conducted onboard trains and surveys mailed to car users in Switzerland, generating a total of 10,728 entries of travel choices among Swissmetro, train, and car. The dataset includes a range of variables such as travel time, cost, travel purpose, ticket type, demographic details of the traveler, and the mode of transport used by the person.

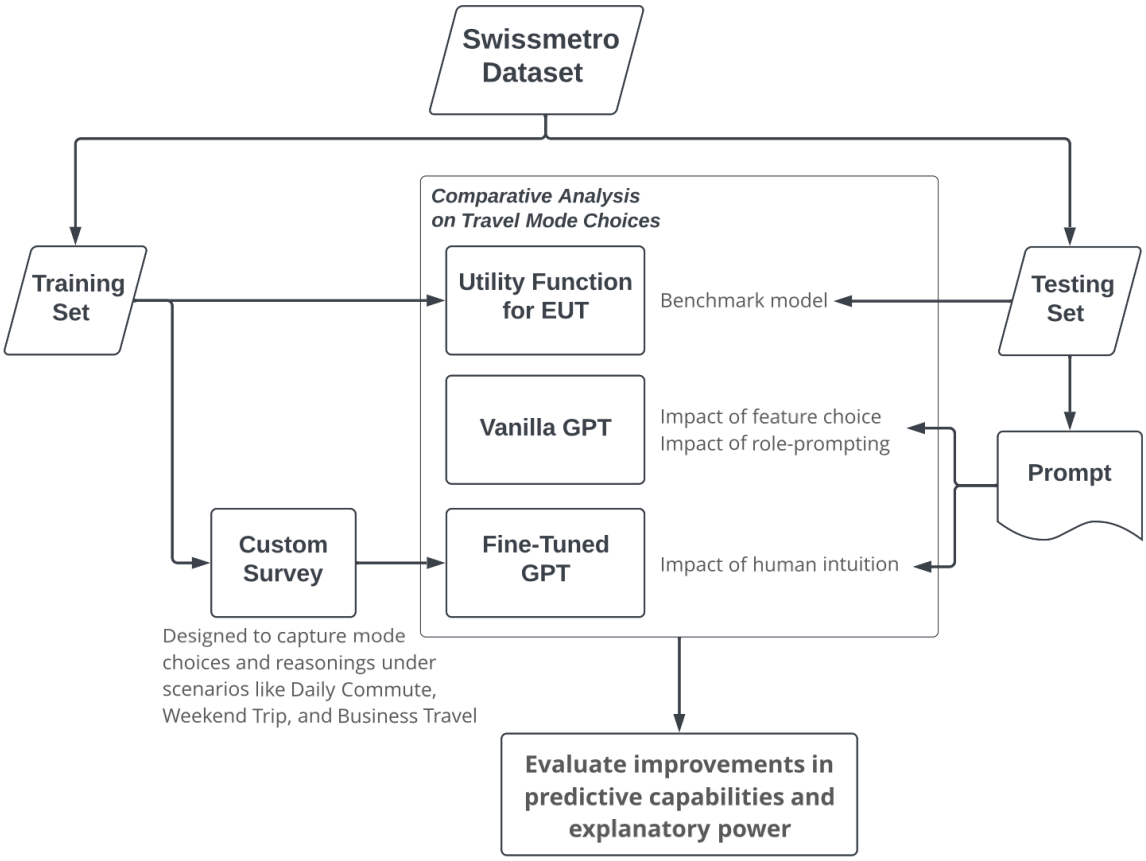


Fig. 1: Method Flow Chart.

B. Testing Against Swissmetro Dataset with the EUT Method

Although Expected Utility Theory works by calculating the expected utility of any given decision and choosing the one with the maximum value, this is difficult to execute in practice because we do not know the exact utilities associated with any outcome. In many cases, we may estimate this using some change in dollar amounts. This works especially well for gambles, where we know exactly how much we stand to lose or gain, but it breaks down when we are dealing with situations such as the one in the SwissMetro dataset. We do not know how much the participants will suffer or benefit from choosing the train, the metro, or their cars. We considered estimating this as the cost of each option, but factors such as time, headway, habits, availability, and others play a significant role in the decision respondents ultimately made. Because of this, we decided to use a multinomial logistic regression model in the R programming language to estimate the relative importance of each factor to the final choice, as well as which alternative it supported the most.

We split the data into training and test datasets, using 70% for the former and the remaining 30% for the latter. We attempted to use the full set of columns included in the original dataset, but removed some due to rank deficiency issues (the train and metro were always available, so they were, for all intents and purposes, the same, non-linearly independent columns). We kept the CHOICE column, representing our dependent variable, as well as columns that were both independent and relevant to the final decisions. These ended up being `GROUP`, `AGE`, `INCOME`, `PURPOSE`, `MALE`, `GA`, `WHO`, `TRAIN_CO`, `CAR_CO`, `SM_CO`, `SM_TT`, `TRAIN_TT`, `CAR_TT`, and `CAR_AV`). Notice that `CAR_AV` (car availability) is kept although `TRAIN_AV` and `SM_AV` were not. This is because many individuals participating in the

survey did not have access to private vehicles, but the train and metro apply to everyone the same. The GA variable represents whether a participant has a Swiss annual season ticket for the metro (1 if they do and 0 if they do not). `WHO` represents who would pay for the transportation, the `_CO` variables are the cost, `_TT` is total time, and `GROUP` is whether they are current drivers or current rail users.

After cleaning the data, we turned the dependent variable, `CHOICE`, into a factor and releveled the table with a `CHOICE` reference of 0. In R, a factor is simply a categorical variable. Due to the fact that the `CHOICE` column contains numbers in the original dataset (0 for unkown, 1 for train, 2 for metro, and 3 for car), it is automatically interpreted as an integer value. Converting it to a factor allows the model to know that it is, in fact, a categorical variable capable of taking on 4 distinct values. We also call the `relevel` function with a reference of "0" to order the alternatives from 0 to 3 in increasing order. We store the releveled alternatives in a new column we call "out".

After our data is cleaned and contains the correct data structures, we import R's `nnet` (neural net) library, useful for feed-forward neural networks and multinomial log-linear models. We create a model using the `multinom` function and provide it with the appropriate parameters. First, we need the dependent column, which in this case is "out", as it contains the choices as factors and is releveled using "0" as the baseline. Then, we need the columns holding our independent variables, GROUP, AGE, and the others described above. The last requirement is data, which, in this case, is the cleaned data from before.

To measure the accuracy of our model, we generate a confusion matrix from the results of calling R's predict function on the cleaned test data. We sum the diagonal components, divide them by the sum of all the entries in the table, and save the result for our accuracy.

Having established a benchmark, we then shifted our focus to exploring the capabilities of LLMs.

C. Testing Against Swissmetro Dataset with the Vanilla GPT

In this phase, we evaluated the Large Language Model, specifically ChatGPT 3.5, referred to as 'Vanilla GPT'. The model's predictions regarding travel mode choices were directly compared with our EUT benchmark.

1) GPT Application Programming Interface (API): To assess the ability of LLMs, the study leveraged Open AI's API of the model GPT-3.5 [11]. Each call requires a textual input specifying the task or question to be answered by the GPT model, which is often called a $prompt$. The API returned the model's generated output. The $temperature$ parameter was set as 0 to avoid variances in the API outputs.

2) Feature Choice: This part of the research analyzed how different features might influence the model output reasonings. Three sets of conditions were experimented with. First, only features related to travel time and travel costs were provided in the background information section. Second, travel time, travel cost, and demographic information of the traveler were included in the background section. The demographic information includes income (high or low income), gender (male or female), and age (senior or not), all binary variables. Lastly, other features provided in the dataset were fed into the prompt, which includes luggage information, Swiss annual season ticket holdings, current mode usage, travel purposes, and who pays for the trip. To be able to compare with the EUT results, the feature choice in the third scenario matches the features used in EUT.

The inferred travel mode and the reasoning provided by the GPT model were recorded and compared with the actual mode choice. The assessment process included comparing the LLM's output accuracy with that of the EUT using the same sets of input features. The reasonings of the LLM output were also compared with the EUT. To do this, we analyzed the occurrence of the feature in the reasoning in each trial of the testing and took the percentages of the occurrence frequency. In this way, we can understand what features are important for LLMs and how they compare with EUT.

3) Prompt Design: A simple prompt template was used, including two sections: background information about the person's trip, which includes the feature choices described above, and a description of the task of making inferences about the travel mode choice. For the fair comparison, we restricted the amount of information given to the GPT agent. For instance, we didn't explicitly tell the language model about the Swissmetro, as we didn't in the community survey. We expect the LLMs to learn about the new travel mode from the different travel costs and time information. The numerical features are converted into sentences so that GPT can understand them better. To analyze how language instruction can influence the prediction result and to elicit the relevant knowledge in the LLM training, we provided two prompts for LLMs, one instructs GPT to think as rational individuals, while the other tells GPT they are influenced by behavioral theories such as Prospect Theory. The Prospect Theory is understood by LLM from two sources. First, we induce its knowledge on the Prospect theory based on the large corpus of texts it has trained on. Second, we explicitly told some main ideas in Prospect theory such as risk aversion. Examples of each prompt are presented below:

> **Rational Agent Prompt:** Consider yourself as a rational human. You are a male older than 65 years old. Taking the Train takes 153 minutes and costs 139 Swiss Franc. Taking the Swissmetro takes 117 minutes and costs 181 Swiss Franc. Taking a Car takes 104 minutes and costs 123 Swiss Franc. You are taking luggage with you. You do not own a Swiss annual season ticket for the rail system and most local public transportation. You are a current car user. You are traveling for business purpose. You and your employer share the travel cost. Will you take the Train, Swissmetro, or Car? Think step by step. Organize your answer in a JSON object with two keys: 'prediction' (the predicted travel mode, one of Train, Swissmetro, or Car) and 'reason' (explanation that supports your inference).
>
> **Prospect Theory Aware Agent Prompt:** Consider yourself as a human influenced by the prospect theory, and want to avoid uncertainty and risks. Thus, you tend to choose your current mode of transportation unless other modes are much superior than the current ones. You are a male older than 65 years old. Taking the Train takes 153 minutes and costs 139 Swiss Franc. Taking the Swissmetro takes 117 minutes and costs 181 Swiss Franc. Taking a Car takes 104 minutes and costs 123 Swiss Franc. You are taking luggage with you. You do not own a Swiss annual season ticket for the rail system and most local public transportation. You are a current car user. You are traveling for business purpose. You and your employer share the travel cost. Will you take the Train, Swissmetro, or Car? Think step by step. Organize your answer in a JSON object with two keys: 'prediction' (the predicted travel mode, one of Train, Swissmetro, or Car) and 'reason' (explanation that supports your inference).

The final step in our methodology was to fine-tune the LLM using custom survey data, aiming to align the model more closely with human reasoning patterns in travel mode choices.

D. Fine-Tuning LLM with Custom Survey Data and Testing Against Swissmetro Dataset

The second part of the study involved enhancing the LLM's capability in predicting travel mode choices through a custom-designed survey and subsequent testing against the Swissmetro dataset.

1) Survey: A custom survey was conducted to gather data on travel mode choices and reasoning across three distinct scenarios: Daily Commute, Weekend Trip, and Business Travel. The survey began by collecting demographic information, including gender, age range, and annual household income, to understand the diverse backgrounds of the respondents. In each scenario, participants were presented with unique travel situations and offered choices between a Car, Train, or SwissMetro, each with specific travel times and costs. Participants were required to provide explanations for their choices, a crucial aspect as these reasonings were used to contextualize and fine-tune the Vanilla GPT model.

2) Fine-Tuning GPT Model: Drawing from the insights gained from the survey, the Vanilla GPT model (ChatGPT 3.5) was fine-tuned to align more closely with human reasoning patterns and preferences in travel mode selection. We customize the GPT model for our application by using fine-tuning

features offered by OpenAI [12]. We put the chat history of all choice results and their reasoning data from the survey in text form. This fine-tuning process involved adjusting the model to incorporate the nuances and specific reasoning highlighted in the survey responses. During the fine-tuning process, we trained 1,190,420 tokens, and it showed 0.1104 training loss. Figure 2 shows the training loss and accuracy of fine-tuning process of GPT models.

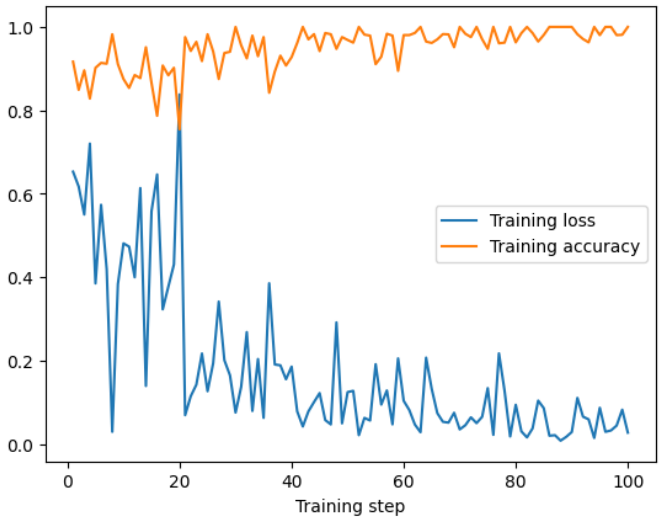


Fig. 2: Training loss and accuracy of fine-tuning GPT models.

3) Testing the Enhanced Model Against Swissmetro Dataset: Once fine-tuned, the enhanced GPT model was tested against the Swissmetro dataset. This step was crucial to evaluate the improvement in the model's performance and its ability to offer better explanatory power for people's travel mode choices. The goal was to determine whether the fine-tuned model could more accurately mimic human decision-making processes and provide reliable predictions in the context of travel mode selection.

E. Evaluation Criteria

Both accuracy and the macro F1 score were used to evaluate the mode choice prediction in LLM and EUT. In addition, reasoning output from LLM was evaluated manually to check if the logic makes sense and how it assigns priority to different features. Since the second step involved manual inspection, for ease of evaluation, this section only used 150 samples, 50 from each travel mode.

V. Results and Interpretation

A. Overall Results

TABLE I: Overall Results

Model Specification	Feature Choice	Accuracy	Macro F1 Score
EUT	travel time and cost + demographic + other	**0.691** $\pm$ 0.000	0.59 $\pm$ 0.000
Vanilla GPT, Rational Agent	travel time and cost	0.456 $\pm$ 0.003	0.382 $\pm$ 0.004
Vanilla GPT, Rational Agent	travel time and cost + demographic	0.493 $\pm$ 0.014	0.430 $\pm$ 0.020
Vanilla GPT, Rational Agent	travel time and cost + demographic + other	0.491 $\pm$ 0.006	0.473 $\pm$ 0.006
Vanilla GPT, PT-Theory Aware Agent	travel time and cost	0.582 $\pm$ 0.011	0.569 $\pm$ 0.010
Vanilla GPT, PT-Theory Aware Agent	travel time and cost + demographic	0.607 $\pm$ 0.011	0.597 $\pm$ 0.012
Vanilla GPT, PT-Theory Aware Agent	travel time and cost + demographic + other	0.631 $\pm$ 0.008	0.620 $\pm$ 0.008
Fine-Tuned GPT, PT-Theory Aware Agent	travel time and cost	0.551 $\pm$ 0.029	0.529 $\pm$ 0.033
Fine-Tuned GPT, PT-Theory Aware Agent	travel time and cost + demographic	0.561 $\pm$ 0.036	0.532 $\pm$ 0.050
Fine-Tuned GPT, PT-Theory Aware Agent	travel time and cost + demographic + other	**0.636** $\pm$ **0.012**	**0.625** $\pm$ **0.004**

Fig. 3: Confusion matrix of results from multinomial expected utility theory model

```
Coefficients:
  (Intercept)     GROUP      AGE    INCOME  PURPOSE     MALE     TRAIN_CO       CAR_CO       SM_CO       SM_TT
1   237.42140 -31.95060 6.372116 -19.21075 8.180143 52.31749 -0.007103416 -0.001509014 -0.01052214 -0.6790579
2   236.64241 -30.92819 6.097899 -19.15622 7.985287 52.68680 -0.005653403  0.003739249 -0.01179652 -0.6867233
3    66.87889 -28.83878 6.223662 -19.15177 8.126048 52.46285 -0.006360896 -0.004479825 -0.01118399 -0.6821051
   TRAIN_TT      CAR_TT    CAR_AV
1 0.5338841 -0.05544963 -56.07550
2 0.5410502 -0.05609047 -55.68919
3 0.5499062 -0.06971013 107.74015

Std. Errors:
  (Intercept)     GROUP       AGE    INCOME   PURPOSE      MALE   TRAIN_CO    CAR_CO       SM_CO       SM_TT
1  0.09113641 0.2759257 0.2108115 0.2449590 0.2505660 0.3205563 0.03113081 0.3133697 0.001376917 0.006247872
2  0.08969533 0.2800437 0.1807586 0.2018545 0.2126623 0.3501571 0.03402247 0.2988840 0.001121654 0.005239509
3  0.08688126 0.2368817 0.3169989 0.3448100 0.3475846 0.1294095 0.02155339 0.4047583 0.001823396 0.009191248
      TRAIN_TT     CAR_TT      CAR_AV
1 0.0009717526 0.01376668 0.008209131
2 0.0007862913 0.01223726 0.007125765
3 0.0012607080 0.02263438 0.012917775

Residual Deviance: 10842.08
AIC: 10920.08
```

Fig. 4: Coefficients yielded by the multinomial logistic regression model

B. Expected Utility Theory

Running the predict function using the multinomial EUT model yielded the confusion matrix in Figure 3. For the test on the full dataset, using 70% of the data to train and 30% to test, dividing the diagonal elements of the confusion matrix by the total number of predictions yielded $\frac{2253}{3259} = 0.691$. Our EUT model, not augmented with machine learning, is able to accurately predict the mode choice individual's will choose in specific situations based on past data over two-thirds of the time. Keeping in mind that we did not, and could not, know any information regarding the individual's independent utility curves, this is [relatively] high result. More sophisticated discrete choice models may have yielded an even higher accuracy.

Although this model is quite accurate given the limited information, there is still room for improvement. As our understanding of LLMs improves, we may be able to take advantage of this new technology to close this gap and have better insight into the mode choices people are likely to make.

Figure 3 shows the coefficients with which our model converged using the multinomial logistic regression approach. The categories with the highest coefficients were GROUP, INCOME, and MALE.

AGE and PURPOSE had a moderate effect on the outcome, with approximately 6 and 8, respectively. Surprisingly, the time, cost, and availability of each of the three modes seemed to all have negligible impact on the choice of the participants. This may imply that demographic information is much more relevant for mode choice prediction than specific information about the current context, especially when that information is not too variable, such as the cost of a train ticket being stable.

C. Vanilla GPT Prediction

As shown in the overall result in Table I, the Vanilla GPT prediction has a relatively low prediction accuracy. Since the share of each mode is equally distributed, randomly guessing each mode would result in an accuracy of around 0.33. The Vanilla GPT is better than that by at least 38%, which demonstrates its prediction is better than the random guess. Moreover, since the prompt is zero-shot, meaning it has not seen any training examples, achieving this level of accuracy means the Vanilla GPT has some common sense and can think analogous to humans to a certain extent.

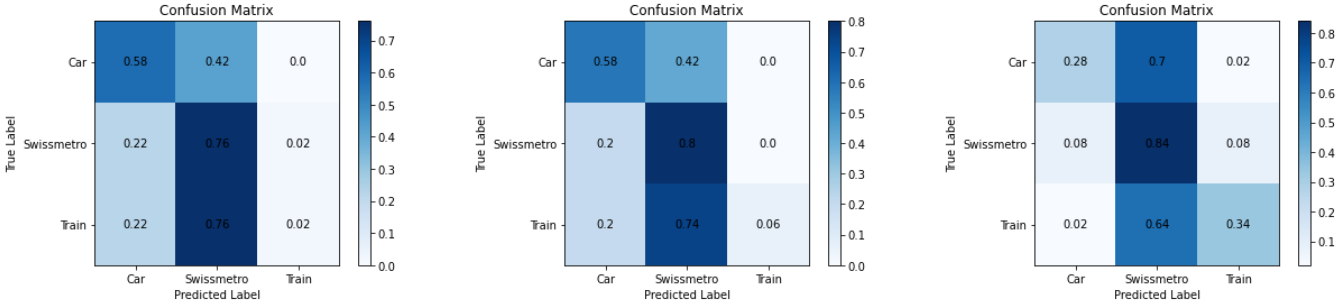
To evaluate how the feature choice influences the prediction outcomes, the accuracy and the macro F1 score are calculated with each combination of the feature choice sets, namely with only travel time and cost, travel time and cost with demographics of the passenger, and travel time and cost with demographics and other traveler information. When we use the rational agent prompt, the accuracy of each set of features is 0.456, 0.493, and 0.491, respectively, while the F1 score is 0.382, 0.430, and 0.473, respectively. In this case, adding more information does not necessarily increase the prediction accuracy but increases the F1 score. This means the additional information can help direct GPT to achieve better predictions in different classes. On the other hand, in the context of the PT-aware agent prompt, adding more features will improve both the output accuracy and the F1 score.

We also demonstrated the effect of designating specific roles in the prompt on the outputs. Specifically, we compared how GPT can make predictions when they assume the role of rational versus PT-aware agents. By changing the role from rational to PT-aware agents, the accuracy improved significantly by 27.6%, 23.1%, and 28.5% for each set of feature choices. Such a dramatic increase could be because when humans make decisions in real life, they do not always act rationally, and by framing the prompt and eliciting the risk-averse aspect of the decision-making behaviors from LLMs, the predictions capture more variances in the data. This also illustrates the powerful impact of role framing in prompt engineering.

To explore the prediction errors in detail, we plotted the confusion matrix for each model specification and feature choice sets, which are shown in Figure 5a, 5b, and 5c. In the rational agent scenario, the model tends to predict Swissmetro when the true label is Train or Swissmetro due to the fact that Swissmetro provides a better level of service most of the time; thus, the rational agent chooses Swissmetro. When we add more features to the prompt, the prediction can better identify the mode choice for Train, but at the same time, the prediction accuracy for Car reduces. This might be because more information can confuse GPT, which factor plays an important role in mode choice decision-making.

On the other hand, when we replace the rational agent prompt with a PT-aware prompt, the confusion matrix improves, as shown in Figure 6a, 6b, and 6c. Specifically, the model can make better predictions when the true label is Car or Train, but its prediction power weakened towards Swissmetro. This could be resulted from the fact that the prompt over-emphasizing the uncertainty and risks associated with the new modes, which is Swissmetro in this case.

Lastly, we analyzed the feature importance in terms of the percentage of each feature occurred in GPT's output reasonings. Figure 7 demonstrates the feature importance for the rational agent in the vanilla GPT. The distribution of the feature importance is very different as displayed by the coefficient in the EUT model. Interestingly, the rational agent is able to figure out that the most important features are 'cost' and 'time', which were considered in almost all the cases. This conforms with human intuition. 'Purpose' also serves an important role, occurring in about 70% of the cases. Other factors such as

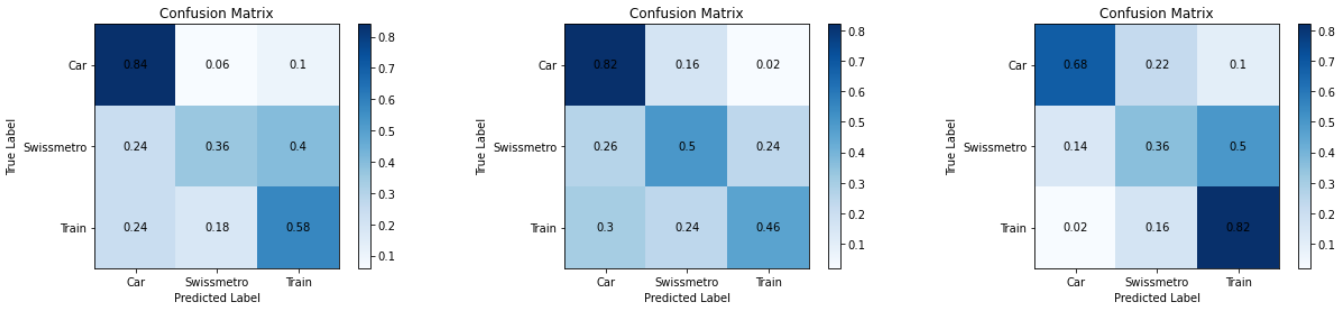


(a) travel time and cost

(b) travel time and cost + demographic

(c) travel time and cost + demographic + other

Fig. 5: Confusion matrix of the vanilla GPT, rational agent prompt



(a) travel time and cost

(b) travel time and cost + demographic

(c) travel time and cost + demographic + other

Fig. 6: Confusion matrix of the vanilla GPT, prospect theory-aware agent prompt

'who' (who pays for the travel), and 'group' (the current mode of transportation that the passenger is using) play a less important role, with less than 30% occurrence in the reasonings.

On the other hand, the feature importance for PT-aware agent in the vanilla GPT has a slightly different focus, which can be illustrated by Figure 8. While the primary features considered are still 'cost' and 'time', the PT-aware agent cares less about the 'purpose' and cares more about 'group' (the current mode of transportation that the passenger is using). This makes sense since the PT-aware agent would consider not only how they are traveling but also what the potential risks associated with that mode are.

To further examine whether the GPT has sufficient math capabilities to come to evaluate the features 'cost' and 'time', we manually examined the first 25 reasoning output by the Vanilla GPT rational agent version with full feature choices. We found out that, in general, GPT has a basic ability to distinguish the magnitude of the values. It is sometimes able to provide the reasoning in terms of the correct time/cost differences between modes. However, it did make several mistakes when comparing the travel time and cost across different modes. In the 25 samples examined, we found three samples with incorrect reasonings, marking the fastest or cheapest mode wrong.

D. Fine-Tuned GPT Prediction

The fine-tuned GPT model, particularly with Prospect Theory Aware Agent configuration, demonstrates a significant improvement in predictive performance. As evidenced in Table I, the model's accuracy and macro F1 score are substantially higher compared to other configurations, especially when all features (travel time and cost, demographic, and other traveler information) are included.

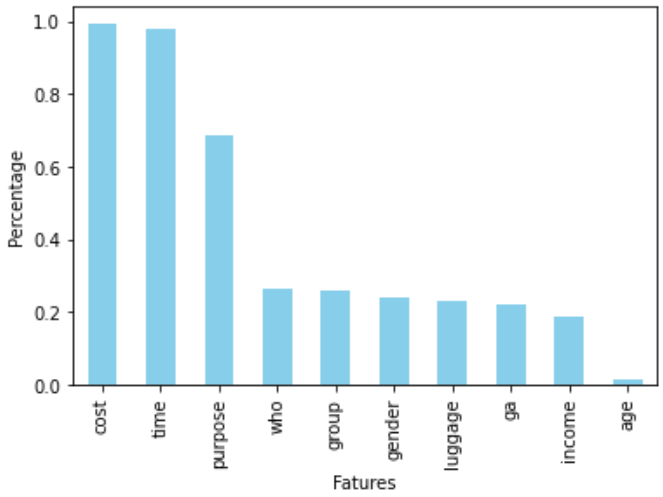


Fig. 7: Percentage of samples utilizing each feature in their reasoning, rational agent prompt with travel time and cost + demographic + other features

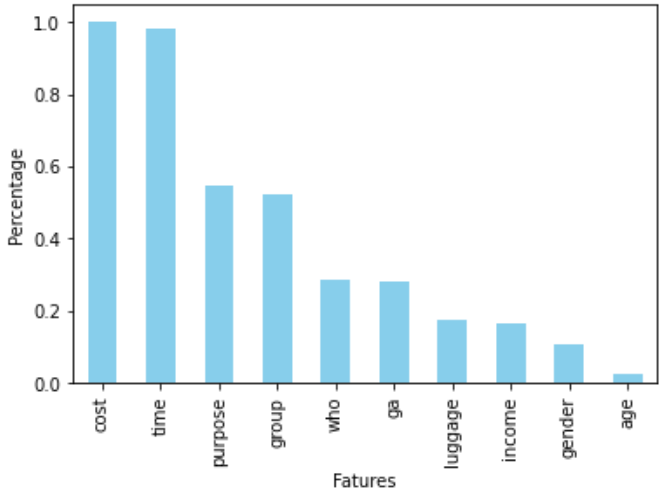
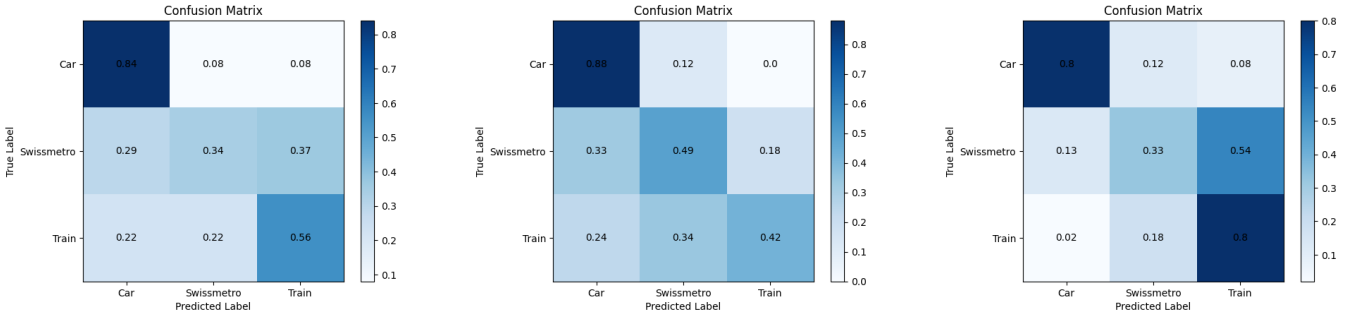


Fig. 8: Percentage of samples utilizing each feature in their reasoning, prospect theory-aware agent prompt with travel time and cost + demographic + other features

When compared with the Vanilla GPT models, the fine-tuned GPT model with all feature sets improves the prediction accuracy. The results indicate that the fine-tuned model benefits greatly from the comprehensive inclusion of features. This fine-tuned model outperforms the other configurations, suggesting that the process of fine-tuning, in conjunction with a broader range of features, optimizes the model's ability to make accurate predictions. The enhanced performance of the fine-tuned model is indicative of its advanced learning and adaptation mechanisms, which are pivotal in handling complex tasks.

Figure 9 displays three confusion matrices for a fine-tuned GPT model with prospect theory awareness, each corresponding to a different set of input features. In Figure 9a, the model strongly favors car predictions with a high true positive rate but is less accurate for Swissmetro and train predictions. Adding demographic data in Figure 9b slightly improves Swissmetro predictions but reduces accuracy for car and train. The most balanced performance across all transport modes is observed in Figure 9c, where the inclusion of travel time, cost, demographic, and other features results in a notable increase



(a) travel time and cost

(b) travel time and cost + demographic

(c) travel time and cost + demographic + other

Fig. 9: Confusion matrix of the fine-tuned GPT, prospect theory-aware agent prompt

in the true positive rate for trains, a moderate rate for Swissmetro, and a slight reduction for cars. This suggests that a richer feature set enhances the model's ability to discriminate between different modes of transportation more effectively.

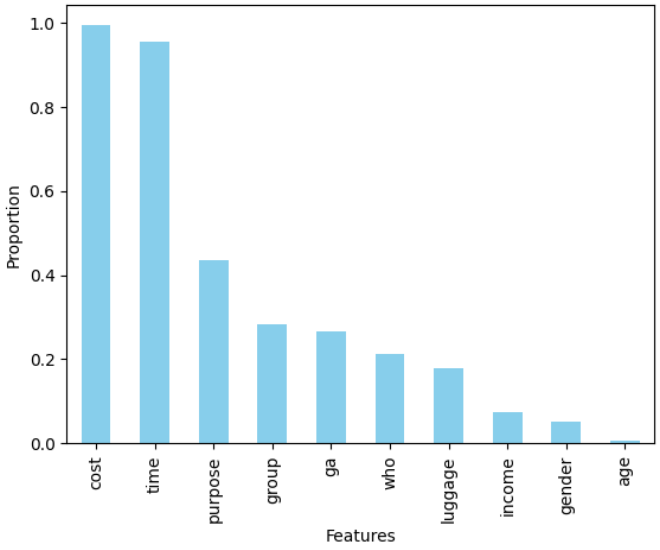


Fig. 10: Percentage of samples utilizing each feature in fine-tuned GPT model's reasoning (prospect theory-aware agent prompt with travel time and cost + demographic + other features)

Figure 10 shows a similar trend with 'cost' and 'time' being the most utilized features in the model's reasoning process. This suggests that regardless of the model configuration, these two features are considered highly influential in the decision-making process of the model. However, there might be differences in the relative importance of subsequent features like 'purpose', 'group', and 'who', which could indicate that fine-tuning the model or incorporating prospect theory influences how the model weights different features beyond the most dominant ones (cost and time).

VI. Discussion

A. Implications and Limitations

This study contributes to the field of urban mobility and AI by highlighting the potential and capabilities of Large Language Models (LLMs) in predicting travel mode choices. The comparison between LLMs and traditional Expected Utility Theory (EUT) models offers a perspective on how

AI can simulate complex human decision-making processes. It also has practical implications for the development of AI-driven mobility solutions, providing insights into the reliability of LLMs in real-world applications. It suggests that LLMs, when fine-tuned with context-specific data, can potentially enhance transportation planning and offer personalized travel recommendations. The study, however, is not without its limitations. The primary limitation lies in the inherent 'black box' nature of LLMs. Despite their advanced capabilities, LLMs often lack transparent reasoning pathways, making it challenging to fully understand and trust their decision-making processes unlike 'logic' models. Furthermore, the study's reliance on the Swissmetro dataset and a specific survey may limit the generalizability of the findings. The biases inherent in the training data of LLMs also pose a risk of skewing the model's outputs, potentially leading to biased or inaccurate predictions.

B. Future Research

Future research could focus on addressing these limitations. Extending the study to include more diverse datasets and real-world scenarios can enhance the explainabilty of LLMs. Meanwhile, further exploration into making LLMs more transparent and interpretable is crucial. Efforts such as mitigating biases in LLMs, ensuring fairness and inclusivity in model outputs, and continually updating models to reflect changing societal norms and behaviors will be vital in advancing the application of AI. Exploring the use of Small Language Models (SLMs) presents a promising avenue as well. SLMs, with their lower computational requirements, offer a more resource-efficient alternative to LLMs, potentially making them more accessible for smaller-scale projects. Additionally, as this study has also examined, SLMs could be fine-tuned to be more specialized for tasks. It would offer more nuanced insights in the context of urban transportation. Finally, the smaller size of SLMs may provide greater transparency and interpretability in their decision-making processes, addressing LLMs' 'black-box' nature.

C. Conclusion

This study demonstrates that LLMs, particularly ChatGPT, are promising in augmenting traditional models for predicting travel mode choices. However, a cautious approach is needed in their application, considering their limitations in transparency and potential biases. As AI continues to evolve, its integration into urban mobility planning must be accompanied by rigorous evaluation and ethical considerations to ensure that it serves as a reliable, fair, and effective tool for enhancing connectivity among people.

APPENDICES

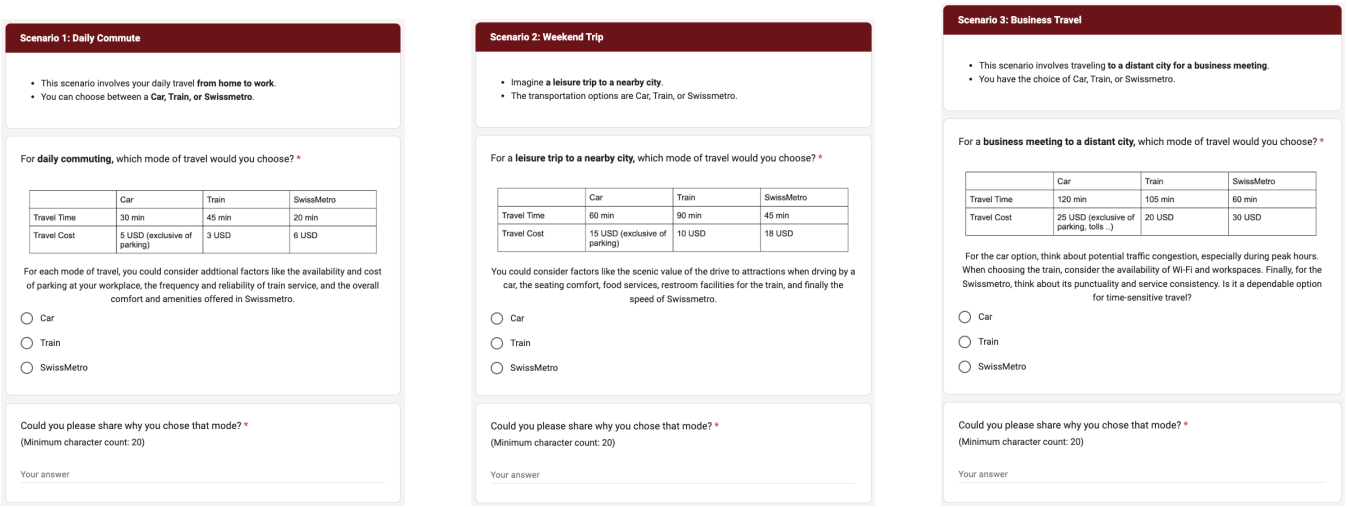
Appendix A: Contributions

- Jung-Hoon Cho: Responsible for fine-tuning the GPT model based on the survey results.
- Youry Moise: Focused on designing the Expected Utility function using the SwissMetro dataset.
- Gigi Sung: In charge of designing the survey and conducting its subsequent analysis.
- Hanyong Xu: Focused on testing and analyzing the reasoning of the Vanilla GPT model using the SwissMetro dataset.

Appendix B: Survey Questions and Results

1) Survey Questionnaire
   - Demographic Information
     - Please select your gender.
       Male; Female; Other/Prefer not to say.
     - Please select your age range.
       <23; 24-39; 40-54; 55-65; 66 or older.
     - Please select your annual household income.
       Less than $56,000; $56,000 to $112,999; $113,000 or higher; Prefer not to answer.

- Overview
  - "In this section, you will be presented with three distinct travel scenarios. Each scenario describes a hypothetical situation with varying travel times, costs, and other parameters. You will be asked to choose your preferred mode of transportation for each scenario and explain the reasoning behind your choice."
  - Three modes of transportation
    - Car
    - Train
    - SwissMetro: an underground transportation network intended to provide a high-speed, efficient, and cutting-edge alternative to the existing forms of public transport in Switzerland.
- Scenario 1: Daily Commute
  "This scenario involves your daily travel from home to work. You can choose between a Car, Train, or Swissmetro." (Figure 11a)
- Scenario 2: Leisure Trip
  "Imagine a leisure trip to a nearby city. The transportation options are Car, Train, or Swissmetro." (Figure 11b)
- Scenario 3: Business Travel
  "This scenario involves traveling to a distant city for a business meeting. You have the choice of Car, Train, or Swissmetro" (Figure 11c)



(a) Scenario 1 (Daily commute) (b) Scenario 2 (Leisure Trip) (c) Scenario 3 (Business trip)

Fig. 11: Survey questions on three different scenarios: Scenario 1 (Daily commute), Scenario 2 (Leisure Trip), and Scenario 3 (Business trip)

2) Survey Results
Figure 12 shows the demographic information of the survey respondents. Figure 13a, 13b, and 13c show the pie chart of choice distribution for daily commute, leisure trip, and business travel, respectively.

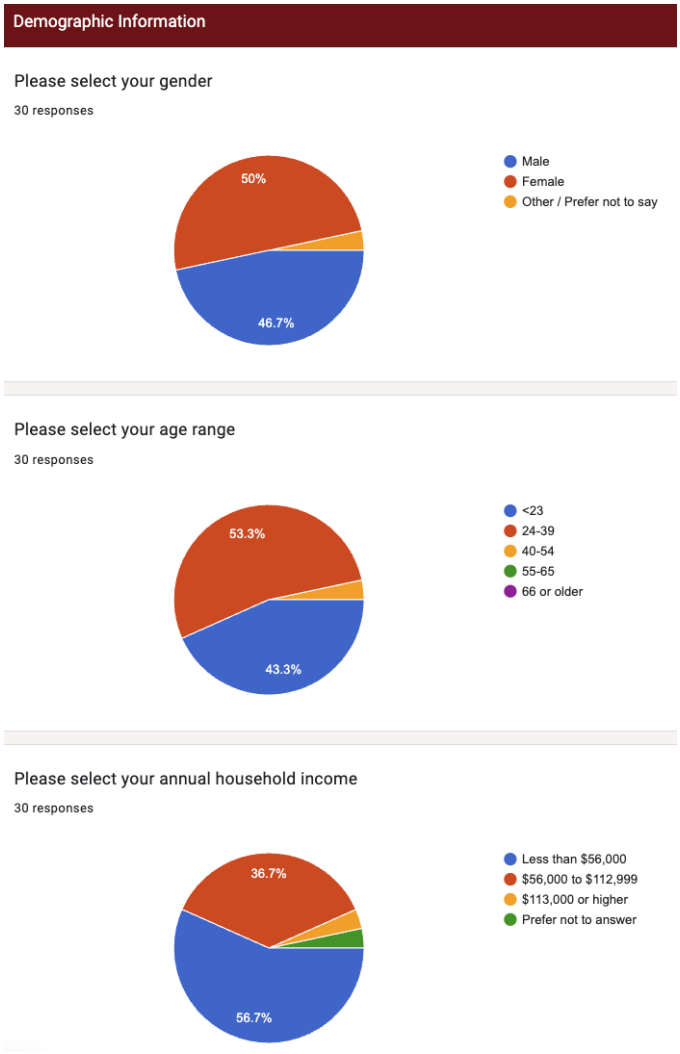
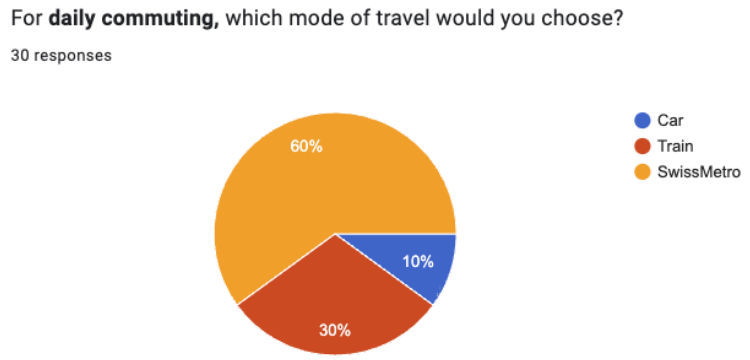
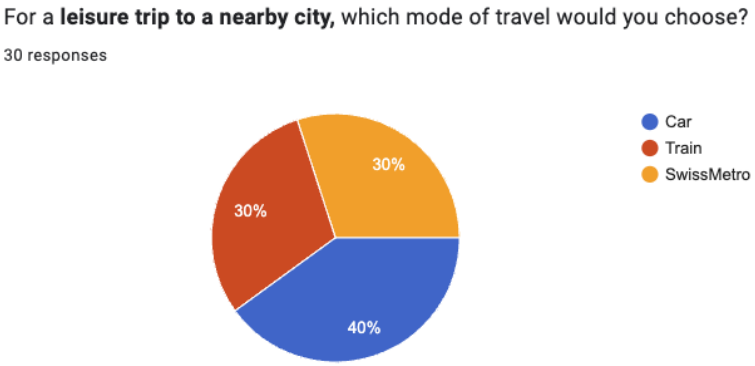


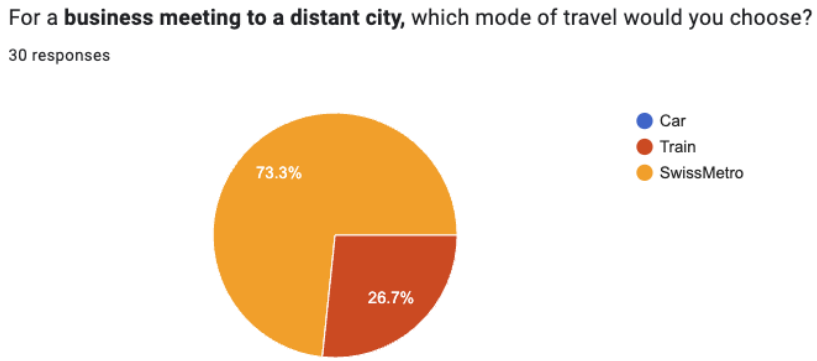
Fig. 12: Survey results: Demographic information



(a) Scenario 1 (Daily Commute)



(b) Scenario 2 (Leisure Trip)

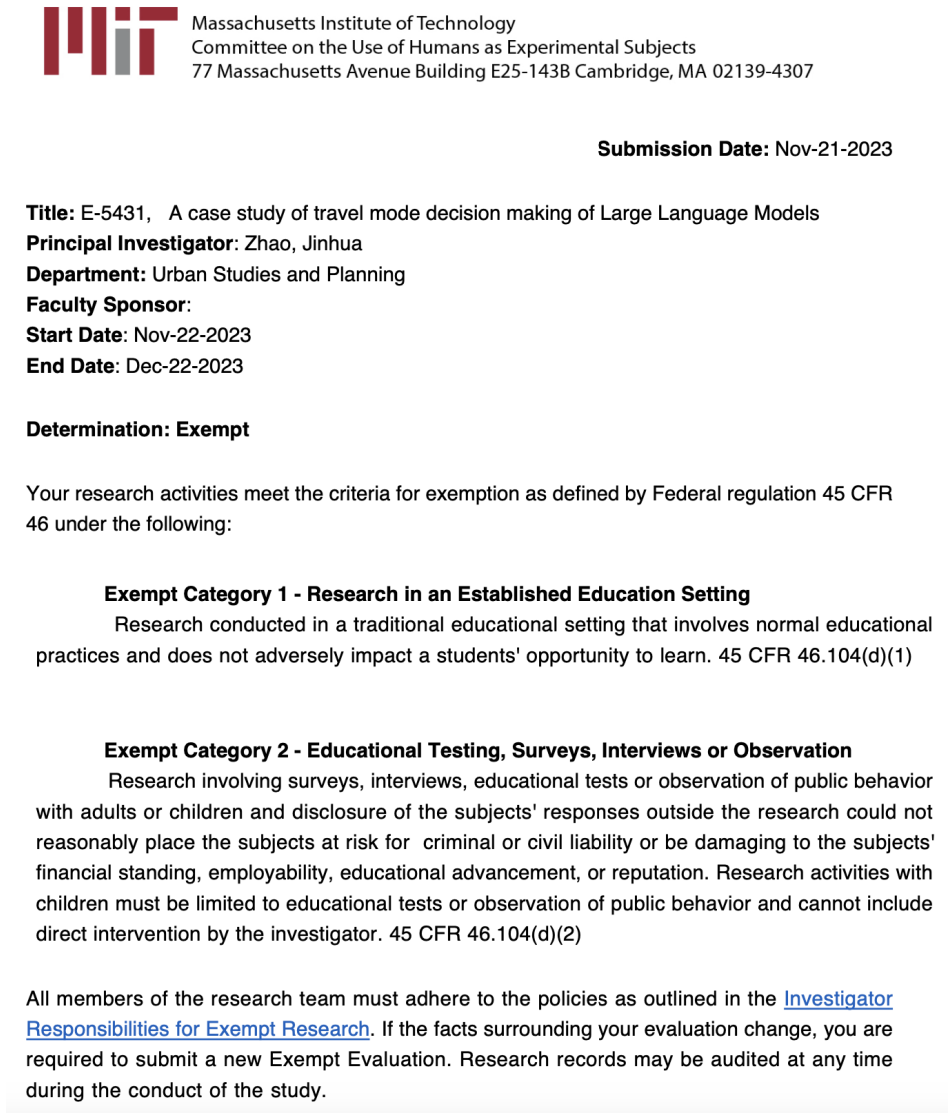


(c) Scenario 3 (Business Travel)

Fig. 13: Survey results on three different scenarios: Scenario 1 (Daily commute), Scenario 2 (Leisure Trip), and Scenario 3 (Business trip)

Appendix C: COUHES Approval

MIT COUHES approved this research as it meets the evaluation exemption criteria. (Figure 14)



Massachusetts Institute of Technology
Committee on the Use of Humans as Experimental Subjects
77 Massachusetts Avenue Building E25-143B Cambridge, MA 02139-4307

Submission Date: Nov-21-2023

Title: E-5431, A case study of travel mode decision making of Large Language Models
Principal Investigator: Zhao, Jinhua
Department: Urban Studies and Planning
Faculty Sponsor:
Start Date: Nov-22-2023
End Date: Dec-22-2023

Determination: Exempt

Your research activities meet the criteria for exemption as defined by Federal regulation 45 CFR 46 under the following:

Exempt Category 1 - Research in an Established Education Setting
Research conducted in a traditional educational setting that involves normal educational practices and does not adversely impact a students' opportunity to learn. 45 CFR 46.104(d)(1)

Exempt Category 2 - Educational Testing, Surveys, Interviews or Observation
Research involving surveys, interviews, educational tests or observation of public behavior with adults or children and disclosure of the subjects' responses outside the research could not reasonably place the subjects at risk for criminal or civil liability or be damaging to the subjects' financial standing, employability, educational advancement, or reputation. Research activities with children must be limited to educational tests or observation of public behavior and cannot include direct intervention by the investigator. 45 CFR 46.104(d)(2)

All members of the research team must adhere to the policies as outlined in the Investigator Responsibilities for Exempt Research. If the facts surrounding your evaluation change, you are required to submit a new Exempt Evaluation. Research records may be audited at any time during the conduct of the study.

Fig. 14: MIT COUHES approval of exempt evaluation

ACKNOWLEDGMENT

The authors express their sincere gratitude for the invaluable support and assistance provided by the teaching team of the *11.478 Behavioral Science, Artificial Intelligence and Urban Mobility* at MIT.

REFERENCES

[1] X. Wang, M. Fang, Z. Zeng, and T. Cheng, "Where Would I Go Next? Large Language Models as Human Mobility Predictors," Aug. 2023, arXiv:2308.15197 [physics]. [Online]. Available: http://arxiv.org/abs/2308.15197

[2] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen, "A Survey of Large Language Models," Sep. 2023, arXiv:2303.18223 [cs]. [Online]. Available: http://arxiv.org/abs/2303.18223

[3] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, "Emergent Abilities of Large Language Models," Oct. 2022, arXiv:2206.07682 [cs]. [Online]. Available: http://arxiv.org/abs/2206.07682

[4] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang, "Sparks of Artificial General Intelligence: Early experiments with GPT-4," Apr. 2023, arXiv:2303.12712 [cs]. [Online]. Available: http://arxiv.org/abs/2303.12712

[5] A. Agrawal, J. Gans, and A. Goldfarb, "How Large Language Models Reflect Human Judgment," Jun. 2023. [Online]. Available: https://hbr.org/2023/06/how-large-language-models-reflect-human-judgment

[6] R. Gordon, "Large language models are biased. Can logic help save them?" Mar. 2023. [Online]. Available: https://techxplore.com/news/2023-03-large-language-biased-logic.html

[7] D. Kahneman, J. Knetsch, and R. Thaler, "Experimental tests of the endowment effect and the coase theorem," *Journal of Political Economy,*, vol. 98, no. 6, pp. 1325–1348, 1990.

[8] D. Kahneman and A. Tversky, "Choices, values, and frames," *American Psychologist*, vol. 39, no. 4, p. 341–350, 1984.

[9] C. K. A. Clayton-Matthews, "Experimental tests of the endowment effect and the coase theorem," *Nursing Research*, vol. 51, no. 6, pp. 404–410, 2002.

[10] M. Bierlaire, K. Axhausen, and G. Abay, "The acceptance of modal innovation: The case of swissmetro," Jan. 2001.

[11] OpenAI, "OpenAI platform." [Online]. Available: https://platform.openai.com

[12] ——, "OpenAI documentation: Fine-tuning." [Online]. Available: https://platform.openai.com/docs/guides/fine-tuning