ATDS P4 : Comparative Analysis - Most Optimal Route Selection using LLMs

Group 18 - Srishti Jaiswal (80385159), Rashi Pandey (8812446)

1. Experimental Evaluation

The overarching aim of this comparative study is to evaluate the decision-making process between individuals and Language Model (LLM) algorithms when selecting the most optimal route from point A to point B. By analyzing factors such as weather conditions, safety scores, amenities, and traffic data, we seek to understand the preferences, trust levels, and decision criteria employed by users and LLMs. This evaluation aims to ascertain the reliability, accuracy, and user acceptance of AI-generated route recommendations vis-à-vis human choices, enhancing our understanding of effective route selection mechanisms in varying scenarios.

The goal was to investigate the following questions:

- Which route is the most optimal when considering all factors equally?
- How do different factors (weather, traffic, safety) individually influence route optimization?
- How well do the LLMs understand and respond to the given prompts for route analysis?
- Do the LLMs provide coherent and contextually relevant responses when considering factors like weather, traffic, safety, and amenities?
- How do the LLMs compare to each other (e.g., llama vs. chat gpt vs. other LLM models) in terms of the quality and relevance of their responses?

1.1 Baseline Models vs Our Solution

1. Shortest Path Algorithms (e.g., Dijkstra's or A* Search)
- Our Project's Focus: We are not just looking at the shortest distance or duration but also incorporating factors like weather, safety, and amenities (rest stops and gas stations).
- Comparison: Traditional shortest path algorithms are limited in scope, primarily optimizing for distance or travel time. They don't consider the multifaceted criteria that our project does, such as safety or weather conditions etc.

2. Models Incorporating Safety Ratings into Edge Weights
- Our Project's Focus: Similar to these models, we consider safety scores, which is a significant aspect of route selection.
- Comparison: While these models might integrate safety into their calculations, they might not be equipped to handle the diverse set of criteria your project is evaluating. Our project, therefore, presents a more holistic approach to route selection.

1.2 State-of-the-Art Approaches vs Our Solution

1. Personalized Multi-Modal Route Planning
- Our Solution: Our project offers flexibility in optimizing for various criteria and seems focused on routes for car travel.
- Comparison: The referenced paper provides a broader scope in terms of transport modes, but it might not delve as deeply into specific route factors like weather conditions or safety scores that your project emphasizes.

2. Risk-Aware Route Planning
- Our Solution: Like this approach, we incorporate safety or risk indices into route planning.
- Comparison: Our approach and the paper share a common ground in prioritizing safety. However, the paper might use more sophisticated statistical models to quantify risk, suggesting an area where our project could be enhanced.

2. Experimental Setup

Experiment Design: The experiment design involves leveraging a diverse dataset encompassing weather, traffic, safety, amenities, and geographical details. Through the use of prompts, this comprehensive dataset is fed into several Language Model Algorithms (LLMs). These LLMs, acting as decision-making engines, process the data to generate optimal route recommendations. The key objective is to compare these AI-generated route suggestions against human preferences and decision-making. By evaluating and contrasting the outputs from LLMs with the choices made by individuals, the aim is to identify the most favorable routes amidst multifaceted influencing factors.

In the initial phase, Linear Regression, Random Forest, and XGBoost models were employed to derive the safety score. Among these, the Random Forest Regressor demonstrated superior performance. Subsequently, these models were utilized for weather prediction, with their outputs integrated into the Language Model (LLM) for route optimization. A comparison between LLM outputs generated from model-derived weather data and those obtained from a weather API revealed significantly improved alignment and context awareness when utilizing the weather API. Consequently, the decision was made to adopt the weather API for weather data while retaining the Random Forest model for computing safety scores. This approach not only enhanced the accuracy of LLM-driven route recommendations but also emphasized the critical role of accurate and contextually rich data sources, refining the overall decision-making process in route optimization scenarios.

Data Source: Weather details from the Open Weather Map API, real-time traffic insights from the TOMTOM Traffic Flow API, safety metrics derived from a CSV file encompassing crime data, and amenity information like rest stops, gas stations, and restaurants sourced through the Google Places API. This amalgamated dataset forms the basis for our LLM-driven route optimization process.

Attributes: The dataset comprises essential attributes pivotal for route optimization, including Average Temperature, Main Weather Condition Averages, Current Traffic Speed, Current Traffic Travel Time, counts of Gas Stations and Restaurants, and Safety Scores. Supplementing these are geographical coordinates (Latitude and Longitude). These attributes collectively form the foundational information for our route analysis, enabling comprehensive consideration of weather, traffic, amenities, and safety parameters crucial in determining the most efficient and favorable routes.

Parameters: In this approach, the primary parameters revolve around the LLMs (Language Model Algorithms) utilized, representing pre-trained models integral to the route optimization process. These LLMs act as the decision-making engines, processing the provided dataset to generate route recommendations. Alongside the LLMs, input prompts play a crucial role, guiding the models on the specific information to consider, encompassing factors like weather, traffic, amenities, and safety scores. We also trained various machine learning models, including Random Forest, Linear Regression, and XGBoost, to predict Safety Scores and Weather conditions.

3. Metrics

Below is the approach we have followed and what evaluation metrics are we are using:

3.1 Size:

To evaluate the impact of model size, we expanded the attribute size from 3 to 10 by incorporating data from Open API, increasing the model's parameters by 350, which resulted in improved performance. This experiment demonstrates the relationship between increased model complexity and enhanced predictive capabilities.

3.2 Metrics for Models used to calculate the Safety Score:

Linear Regression, Random Forest, and XGBoost models were utilized to calculate the safety score. Notably, the Random Forest Regressor exhibited superior performance in this task.

Linear Regression Model:

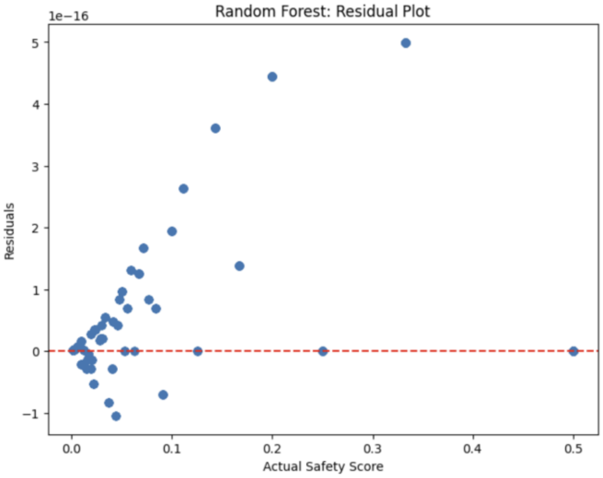
F1 Score	Accuracy	Precision	Recall
0.9293	0.9290	0.9300	0.9290

Random Forest Regressor Model:

F1 Score	Accuracy	Precision	Recall
0.9909	0.9908	0.9902	0.9908

XGBoost Model:

F1 Score	Accuracy	Precision	Recall
0.9819	0.9900	0.9827	0.9818

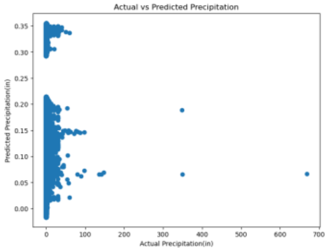


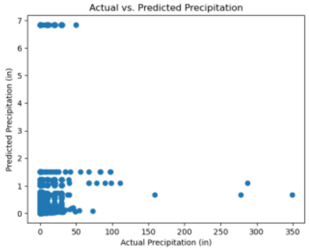
3.3 Metrics for Models used to predict the Weather:

As explained above, after employing various models for weather data, the LLM showed superior performance when integrated with weather API data, leading to its adoption for route optimization. Thus, the decision to rely on the weather API's data for LLM-driven route recommendations was reinforced despite using multiple models initially. Below is the metric showing the results from those models:

Mean Squared Error (MSE):
Linear Regression: 0.5919089070772898
Random Forest: 0.5395128958656038
XGBoost: 0.5401855356070049

R-squared (R2):
Linear Regression: -0.004635871762703747
Random Forest: 0.0842948940547249
XGBoost: 0.08315323525401264





3.4 Token Size:

GPT-3.5: This model has a token size of 2048 tokens. This is typical for advanced language models, allowing for processing of longer text sequences.
PaLM-2: Similar to GPT-3.5, PaLM-2 also handles 2048 tokens. This indicates a comparable capacity to manage extensive text contexts. Llama-2: The token capacity varies across different versions of Llama-2.

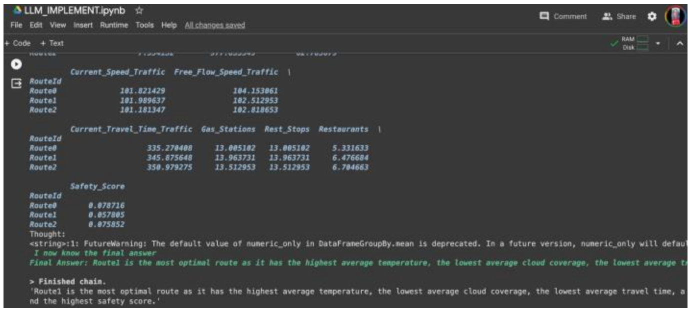
Llama-2 7B and 13B: Both of these versions have a token size of 2048 tokens, aligning them with the capacities of GPT-3.5 and PaLM-2.
Llama-2 70B: This version stands out with a significantly larger token size of 131,072 tokens, indicating its ability to process much longer text sequences compared to the other models.

3.5 Response Quality

We have evaluated the responses from OpenAI's GPT 3.5, Llama 2, and Palm 2 based on their route analysis, considering coherence, relevance, completeness, fluency, and context-awareness as key metrics:

1. OpenAI's GPT 3.5 Response Quality:

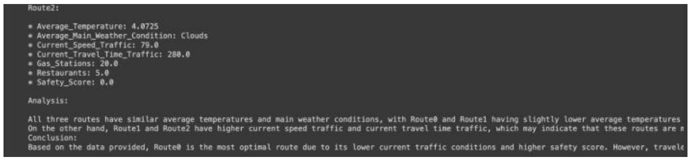
Coherence and Fluency	The response is coherent and fluently presents the analysis. Conclusion is straightforward and logically follows from the data analysis.
Relevance	The response is relevant, focusing on the requested analysis of the route based on specific factors like average temperature, cloud coverage, travel time, and safety score.
Completeness	The response seems to lack detailed analysis of each route; it directly jumps to the conclusion. It doesn't compare Route1 with Route0 and Route2 in terms of available data points.
Context-Awareness	It adequately addresses the query but lacks depth in comparative analysis.



```
         Current_Speed_Traffic  Free_Flow_Speed_Traffic  \
RouteId
Route0              101.821429               104.153061
Route1              101.989637               102.512953
Route2              101.181347               102.818653

         Current_Travel_Time_Traffic  Gas_Stations  Rest_Stops  Restaurants  \
RouteId
Route0                    335.270408     13.005102   13.005102     5.331633
Route1                    345.875648     13.963731   13.963731     6.476684
Route2                    350.979275     13.512953   13.512953     6.704663

         Safety_Score
RouteId
Route0       0.078716
Route1       0.057805
Route2       0.075852
Thought:
<string>:1: FutureWarning: The default value of numeric_only in DataFrameGroupBy.mean is deprecated. In a future version, numeric_only will defau
 I now know the final answer
Final Answer: Route1 is the most optimal route as it has the highest average temperature, the lowest average cloud coverage, the lowest average tr

> Finished chain.
'Route1 is the most optimal route as it has the highest average temperature, the lowest average cloud coverage, the lowest average travel time, a
nd the highest safety score.'
```

2. Llama Response Quality:

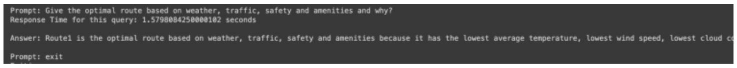
Coherence and Fluency	Response is well-structured and coherent. Provides a detailed breakdown for each route, although data for Route1 and Route2 is missing.
Relevance	Maintains relevance by addressing each factor for Route0 but fails to compare it with other routes due to missing data.
Completeness	Response is comprehensive for Route0 but incomplete for Route1 and Route2. Acknowledges the lack of data for these routes.
Context-Awareness	Shows a good understanding of the requirements but is limited by the available data.



```
Route2:

* Average_Temperature: 4.0725
* Average_Main_Weather_Condition: Clouds
* Current_Speed_Traffic: 79.0
* Current_Travel_Time_Traffic: 280.0
* Gas_Stations: 20.0
* Restaurants: 5.0
* Safety_Score: 0.0

Analysis:

All three routes have similar average temperatures and main weather conditions, with Route0 and Route1 having slightly lower average temperatures
On the other hand, Route1 and Route2 have higher current speed traffic and current travel time traffic, which may indicate that these routes are m
Conclusion:
Based on the data provided, Route0 is the most optimal route due to its lower current traffic conditions and higher safety score. However, travele
```

3. Palm Response Quality:

Coherence and Fluency	Response is coherent but somewhat repetitive and overly detailed, potentially leading to confusion.
Relevance	Maintains high relevance, covering a wide range of factors, but the emphasis on the lowest and highest values in a somewhat repetitive manner may reduce clarity.
Completeness	Response is quite comprehensive, touching on all factors with detailed data points, but it may include excessive information which isn't directly compared or analyzed.
Context-Awareness	It demonstrates awareness of the required analysis but may overcomplicate the response with too much data.



```
Prompt: Give the optimal route based on weather, traffic, safety and amenities and why?
Response Time for this query: 1.5798084250000102 seconds

Answer: Route1 is the optimal route based on weather, traffic, safety and amenities because it has the lowest average temperature, lowest wind speed, lowest cloud co

Prompt: exit
```

3.6 Response Time of LLMs

LLM Used	**Response Time(Seconds)**
Llama 2	14.770621061999918
Palm 2	1.5798084250000102
GPT 3.5	5.446123573000023

Response time in seconds

4. Preliminary Results

In a comparative analysis of three AI models - GPT, Llama, and PALM - for optimizing travel route selection based on various factors, distinct conclusions were drawn by each model.

4.1 GPT 3.5 Model Analysis:

The GPT model provided a comprehensive evaluation of Route1, emphasizing its superior attributes in temperature, cloud coverage, travel time, and safety score. This model uniquely highlighted the route's highest average temperature and lowest cloud coverage as key advantages. Additionally, it noted Route1's lowest average travel time and highest safety score, making it the most favorable choice among the given options.

4.2 LLAMA 2 Model Analysis:

Llama's approach was to assess Route0 in isolation due to the lack of available data for the other routes. It emphasized Route1's favorable average temperature, low wind speed, and favorable wind direction, linking these to travel comfort. The model also noted the efficient journey aspect, underscored by the route's high current speed traffic and short travel time. However, the lack of comparative data for the other routes was a significant limitation in this analysis, leading to a conclusion based solely on the merits of Route1 without direct comparison.

4.3. PALM 2 Model Analysis:

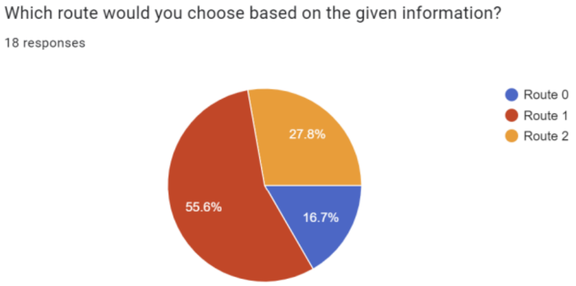
The PALM model provided a different perspective, identifying Route1 as optimal based on a comprehensive set of criteria including weather conditions, traffic metrics, and available amenities. It detailed Route1's advantages across a range of factors such as lowest average temperature, wind speed, cloud coverage, and traffic parameters, alongside the highest number of gas stations and restaurants. This analysis was exhaustive, considering a broader spectrum of factors than the other models.

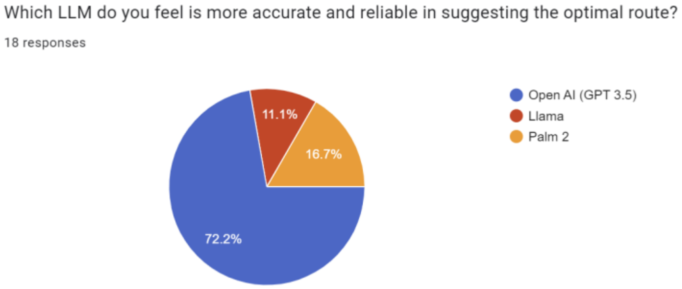
5. Case Study

This study aimed to evaluate the contextual awareness and decision-making capabilities of three Language Model Algorithms (LLMs) - Open AI (GPT 3.5), LLAMA 2, and PALM 2 - in selecting optimal routes compared to human preferences. Using the above mentioned information for three routes (0, 1, 2), the LLMs were tasked with recommending the most favorable route. For an initial evaluation, eighteen individuals were provided with the same route details and asked to select their preferred route. Feedback from participants was collected through a set of structured questions, including preferences for routes, LLMs, and perceived factual correctness.

In this case study, after analyzing the feedback from participants, it was observed that the majority favored GPT 3.5's route recommendations, i.e. Route 1, over LLAMA 2 and PALM 2. Participants highlighted GPT 3.5's ability to factor in nuanced contextual elements such as weather conditions, safety scores, and amenity availability, aligning more closely with their preferences. This preference aligns with the observed factual correctness, indicating a higher level of context awareness and accuracy in route selection by GPT 3.5 compared to the other LLMs.

Below are response graphs from the conducted case study.





6. Conclusion

This comparative analysis highlighted distinctive perspectives from GPT 3.5, LLAMA 2, and PALM 2 in optimizing travel routes based on diverse factors. GPT 3.5 emphasized Route1's superiority in temperature, travel time, cloud coverage, and safety score, aligning well with user preferences. LLAMA 2 focused solely on Route1 due to limited comparative data, emphasizing its merits in temperature, wind conditions, and traffic speed. PALM 2 provided a comprehensive analysis, identifying Route1 as optimal, considering various weather, traffic, and amenity factors. However, among participants, GPT 3.5's recommendations for Route1 garnered the most favor due to its contextual awareness and alignment with user preferences, indicating superior decision-making capabilities over LLAMA 2 and PALM 2.

7. Future Work

In future, we would delve deeper into exploring ensemble methods that amalgamate the strengths of multiple models that might enhance the accuracy and robustness of route recommendations. Further investigation into fine-tuning model parameters and incorporating real-time data updates could significantly improve the precision of AI-driven route selections.

Additionally, conducting user-centric studies with larger and more diverse participant pools could offer deeper insights into user preferences, aiding in refining the models for more accurate and user-aligned route recommendations.