Travel Mode Detection using LLMs

Eklavya Jain
M.S. Data Science, Columbia University
ej2487@columbia.edu

December 19, 2023

Abstract

This paper builds on the work done by [1] using more advanced artificial intelligence algorithms like Large-language models. I experiment with various prompting techniques like persona prompting, chain-of-thought, and in-context learning. Models like Llama2, Mistral, and OpenAI were used to make predictions. A comparative analysis is presented that allows the research team to compare various models through the samples. Human evaluation yielded that Llama2 and Mistral cover for each other's mistakes.

1 Introduction

This paper builds on the work done by [1] using more advanced artificial intelligence algorithms like Large-language models. In [1], the authors leverage Twitter data to understand travel mode choices during the pandemic. This classification, in conjunction with sentiment analysis, is deemed necessary to understand people's attitudinal changes about mode choices during the pandemic. They find that a majority of people had a positive attitude toward buses, bikes, and private vehicles, which is consistent with the phenomenon of many commuters shifting away from subways to buses, bikes, and private vehicles during the pandemic. They further analyzed negative tweets related to travel modes and found that people were worried about those who did not wear masks on subways and buses.

This paper refines and unifies the process of travel mode detection and sentiment analysis using the power of LLMs. I use llama-based models Llama2-7b, and Mistral-7b, and then verify a fraction of the responses using GPT-3.5-Turbo. Various prompt-engineering methods like in-context learning, and chain-of-thought learning are experimented with. The response quality is further enhanced using demonstrations, i.e., few-shot learning. Finally, I align the responses of Mistral and Llama2 and provide a comparative analysis of the responses on the given set of tweets.

2 Literature Review

The base paper [1] uses Twitter data to understand transportation trends before and after COVID-19 hit. They take a two-fold approach to detecting travel modes and then the sentiment of tweets. The authors used BERT along with a classification layer to build the travel mode classifier. They further use BERT with sentiment classification neural network to classify sentiments. This paper builds a good baseline for a world driven by LLMs.

[2] discusses why in-context learning works and explains the role of demonstrations (input-label examples) in in-context learning. The authors analyze how in-context learning performs better than zero-shot inference. Their studies show good results only for classification tasks, which works in our use case of travel mode detection. Since in-context learning allows us to use LLMs in ways without retraining the base model on custom data, it is very efficient and has very little turn-around time. We can nudge it to find answers by using examples and techniques like CoT. This can be very helpful to us as we want to answer specific questions (whose answers are formatted in a specific manner) from reading Twitter posts or Reddit posts.

Further, chain-of-thought prompting [3] allows us to delve into the thought process behind the LLM's generated output. This technique mainly helps with arithmetic reasoning and common sense questions and has only been experimented to work well with models of 100 billion parameters. This technique doesn't seem like a good fit since we use a smaller model and ask textual questions. The one thing that did come out of this paper was the effect of the placement of few-shot exemplars on the generated output.

3 Summary

Code can be found here.
The **results** and the weekly reports can be found here.

4 Methodology

Using various prompting techniques, I asked the LLM to generate

1. the mode of travel
2. the underlying sentiment
3. and the reasoning for the detected sentiment

for a given tweet.

I experiment with persona prompting, followed by chain-of-thought learning, and then finally in-context learning. I also experiment with multiple prompt styles, and variations of the questions to get the most relevant and usable[1] answers. After experimenting, I found that few-shot learning yields the best output, both in terms of accuracy and output format.

In order to verify the answers generated by Llama2, I used another open-source model Mistral-7B to answer the same three questions based on the given tweets. These models are loaded into memory through the bitsandbytes[2] package that quantizes the model and allows faster computation. I, then do a comparative chart analysis of the mode of travel detected, and the sentiment of the tweets. In addition, I get alignment (similarity) scores between the reasoning generated by Llama2 and Mistral to understand and verify the outputs generated by the two models.

Some key points about the Llama2 model used for inference are as follows:

[1] An output format that can be easily parsed and converted to a table format

[2] Source of BitsandBytes

- Tokenizer used: BPE model based on SentencePiece.
- It is an encoder-decoder architecture
- Uses grouped-query attention in addition to the attention masks uses in Llama
- Llama2 was trained using 2 objectives
  - LLM: Autoregressive, i.e., predict next word using the history
  - RLHF: Reward model such that the chosen response has a higher score than the rejected response (binary ranking loss)
- Takes around 10 seconds to run inference on each example on an L4 GPU with 24GB memory and temperature set as 0.01.

Going one step further, I use OpenAI's GPT-3.5-Turbo on 60% of the samples and conduct a similar comparative analysis between the three models. Finally, I perform manual human verification on the outputs produced by both Llama2 and Mistral.

5 Experiments and Results

I used the HuggingFace API to use the Llama2-7B-Chat-HF model by Meta. Initially, I created a basic text generation pipeline with a persona prompting technique. Figure 1 shows the kind of responses that are generated for three sample tweets.
This technique did not work well for three reasons

- It did not detect the travel modes accurately on the sample queries.
- It was unable to segregate 'unrelated' tweets
- The format of the response was unpredictable and unusable.

To eliminate these issues, a thorough exploration of various prompting techniques and query variations was done. I did the following experiments

1. Instead of asking the sentiment, ask whether the user is satisfied or dissatisfied with the service.
2. Chain-of-thought to incorporate tweets that are not related to any travel mode.
3. Chain-of-thought to explain how the LLM is making a prediction.

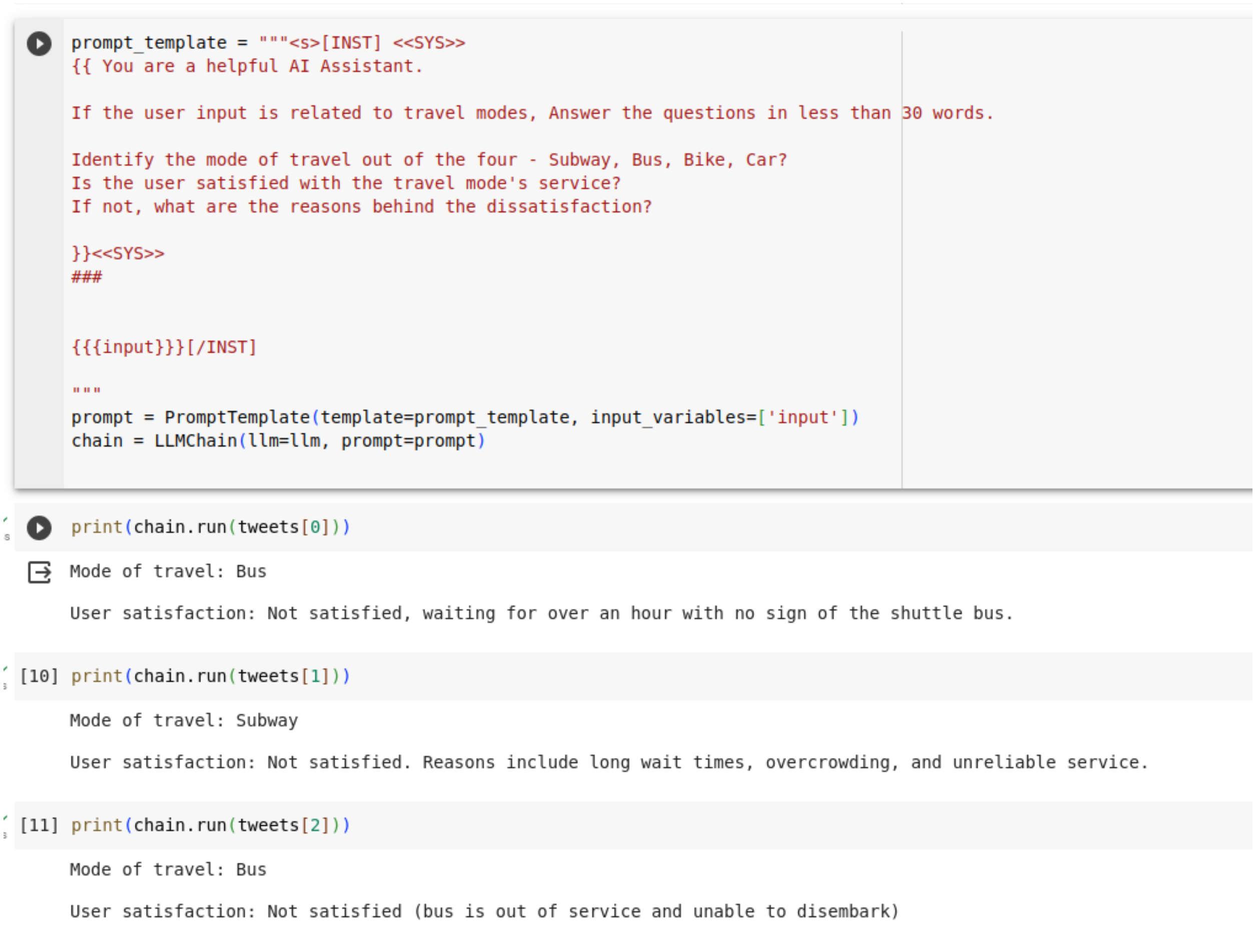
The various prompts that were used to generate predictions are:

- Prompt with different questions (satisfied/dissatisfied):

```
"""
<s>[INST] <<SYS>> You are a natural language expert good at analyzing human
   sentiment toward travel modes by reading text. <</SYS>>
Input: {input}

Is the given input related to travel modes?

If yes, answer the following questions:
```



```
prompt_template = """<s>[INST] <<SYS>>
{{ You are a helpful AI Assistant.

If the user input is related to travel modes, Answer the questions in less than 30 words.

Identify the mode of travel out of the four - Subway, Bus, Bike, Car?
Is the user satisfied with the travel mode's service?
If not, what are the reasons behind the dissatisfaction?

}}<<SYS>>
###


{{{input}}}[/INST]

"""
prompt = PromptTemplate(template=prompt_template, input_variables=['input'])
chain = LLMChain(llm=llm, prompt=prompt)
```

```
print(chain.run(tweets[0]))
```

```
Mode of travel: Bus

User satisfaction: Not satisfied, waiting for over an hour with no sign of the shuttle bus.
```

```
[10] print(chain.run(tweets[1]))
```

```
Mode of travel: Subway

User satisfaction: Not satisfied. Reasons include long wait times, overcrowding, and unreliable service.
```

```
[11] print(chain.run(tweets[2]))
```

```
Mode of travel: Bus

User satisfaction: Not satisfied (bus is out of service and unable to disembark)
```

Figure 1: Llama2 inference using Persona prompting technique

```
1. Identify the mode of travel out of the four - Subway, Bus, Bike, Car?
    Answer in one word.
2. Is the user satisfied with the travel mode’s service? Answer in yes/no.
3. What are the reasons behind the satisfaction/dissatisfaction? Answer in
    less than 20 words.

If no, answer by saying the word ‘unrelated’.
[/INST]

Output: {output}
</s>
"""
```

- Prompt with explicit ask - class defined for unrelated travel mode

```
"""
<s>[INST] <<SYS>> You are a natural language expert who is good at analyzing
    human sentiment toward travel modes by reading text. <</SYS>>
Input: {input}

1. Is the given input related to travel modes?
```

```
2. Identify the mode of travel out of the five if the previous answer is yes -
    Subway, Bus, Bike, Taxi, Car? Answer in one word.
3. What is the sentiment (positive/neutral/negative) of the input? Answer in
   one word.
4. What are the reasons behind the sentiment? Answer in less than 20 words.
Don't give any other explanations.
[/INST]

Output: {output}
</s>
"""
```

- Prompt with the implicit ask - is the input related to travel modes?

```
"""
<s>[INST] <<SYS>> You are a natural language expert good at analyzing human
   sentiment toward travel modes by reading text. <</SYS>>
Input: {input}

answer = Is the given input related to travel modes?
if answer is yes:
Respond to the following three questions:

1. Identify the mode of travel out of the five if the previous answer is yes -
    Subway, Bus, Bike, Taxi, Car? Answer in one word.
2. What is the sentiment (positive/neutral/negative) of the input? Answer in
   one word.
3. What are the reasons behind the sentiment? Answer in less than 20 words.

if answer is no:
Just say that the input is unrelated to travel modes. Don't give any other
   explanations.
[/INST]

Output: {output}
</s>
"""
```

- Chain of thought prompt:

```
"""
<s>[INST] <<SYS>> You are an expert at understanding text, identifying travel
   modes and human sentiment toward travel modes.<</SYS>>
Input: {input}

Is the given input related to travel modes?
Respond to the following three questions and give an explanation for your
   conclusion:

1. Identify the mode of travel out of the five- Subway, Bus, Bike, Taxi, Car?
   Answer in one word.
2. What is the sentiment (positive/neutral/negative) of the input? Answer in
   one word.
3. What are the reasons behind the sentiment? Answer in less than 20 words.

Otherwise, just say that the input is unrelated to travel modes.
[/INST]

Output: {output}
```

```
</s>
"""
```

These prompts are tested with and without any demonstrations and its observed that demonstrations enable the LLM to learn the expected format of the output. The summary of these experiments is presented in Table 1.

Number of Exemplars	Time Taken for 3 Sample Queries (sec)	Queries with Expected Output (out of 3)	Type of Questions	Special Instructions / Modifications
4	33.4	3	Satisfied / Dissatisfied	Don't give explanations for unrelated inputs.
5	40.5	2	Sentiment	Don't give explanations for unrelated inputs.
5	42.2	3	Sentiment	Don't give any explanations.
6	42.34	2	Sentiment	Changed few-shot output pattern from the list to comma-separated values
6	54.03	1	Sentiment + Is related	Converted 3 questions into 4 questions (1 related/unrelated to travel modes)
6	34	1	Sentiment + Is related	Asking if the input is related to the first question and continue asking others
6	52	2	Sentiment + Is related	Updated few-shot exemplars to ignore sentiment questions on unrelated inputs
6	36.78	2	Sentiment	Implicitly asking if the input is related to travel mode or not
4	61.05	3	Sentiment	Chain of thought & Few-shot

Table 1: Experiment Results

Both Llama2 and Mistral have an expected input format using the SYS and INST tokens. The demonstrations were added considering these tokens. An example prompt is shown below

```
<s><<SYS>> You are a natural language expert <</SYS>>
[INST]
Tweet: The subway is delayed yet again. This city just can not run on time.
    Apparently there is a water leak near Times Square. Feel so angry.
```

```
Question: Only answer the following questions in order as bullet points.
1. Select the mode of travel: Subway, Bus, Bike, Taxi, Car, Unknown
2. Select the sentiment: Positive, Neutral, Negative
3. Explain your reasoning behind the selected sentiment in less than 20 words.
[/INST]

Answer:
1. Subway
2. Negative
3. The subway was delayed due to a water leak near Times Square.
</s>
```

Consequently, a prompt template was generated for making predictions.

```
<s><<SYS>> You are a natural language expert <</SYS>>
[INST]
Tweet: {input}

Question: Only answer the following questions in order as bullet points.
1. Select the mode of travel: Subway, Bus, Bike, Taxi, Car, Unknown
2. Select the sentiment: Positive, Neutral, Negative
3. Explain your reasoning behind the selected sentiment in less than 20 words.
[/INST]

Answer:
</s>
```

This prompt supported by 4 demonstrations was able to generate accurate and well-formatted responses. A GCP instance with an L4 GPU with 24GB memory was set up to run inference on all the 2000 samples.

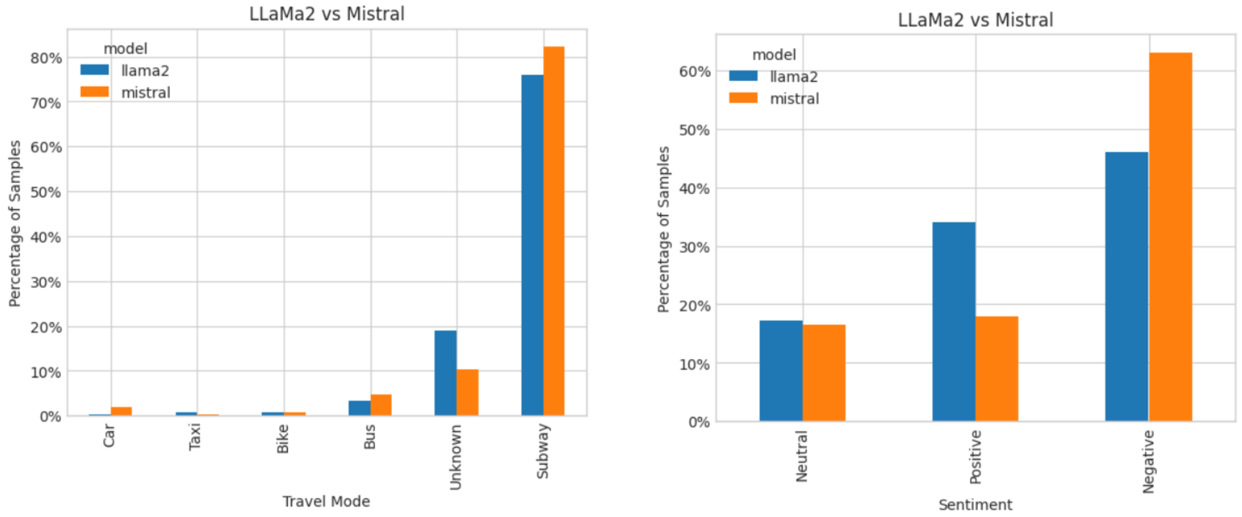
6 Results and Comparisons

Figure 2 shows the distribution of the detected mode of travel and tweet sentiment across all the 2000 samples. I observed that most of the samples were detected as tweets related to Subway followed by unrelated ones. Darren validated this by sharing more details on the subset of samples shared with me.

While the above chart gives us a fair idea of how the overall predictions fair between the two models, there is no information about whether these predictions align pairwise. Figure 3 shows the pairwise alignment of the detected mode of travel and the sentiment.

We observe that for the mode of travel, the majority of samples either align perfectly or misalign with the *Unknown* category. On careful human evaluation, I observed that a major fraction of the misaligned samples, i.e., *(Unknown, Subway)* and *(Subway, Unknown)*, can be labeled as Subway as both Llama2 and Mistral can cover for each other's faults.

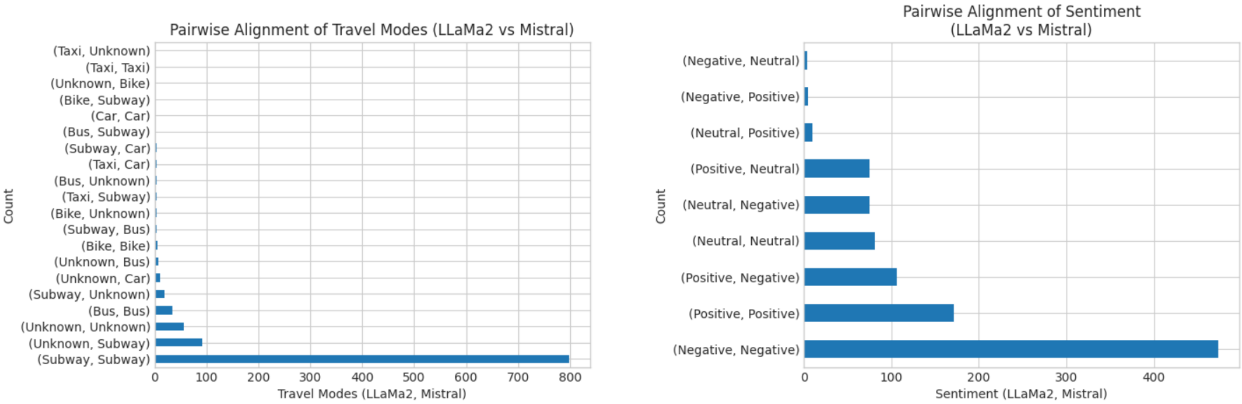
Further, I used OpenAI's API to run inference on 1200 samples. Figure 4 compares the overall numbers for mode of travel and sentiment produced by all the models. Observe that there is only a minor change in the percentages of samples that are classified as *Subway*. This just goes on to show that open-source LLMs are equally good, if not better at predicting the underlying mode of travel and sentiment of tweets. A total of $20 was spent on the OpenAI API.



(a) Mode of Travel

(b) Tweet Sentiment

Figure 2: Llama2 vs Mistral on all 2000 samples



(a) Mode of Travel Alignment

(b) Tweet Sentiment Alignment

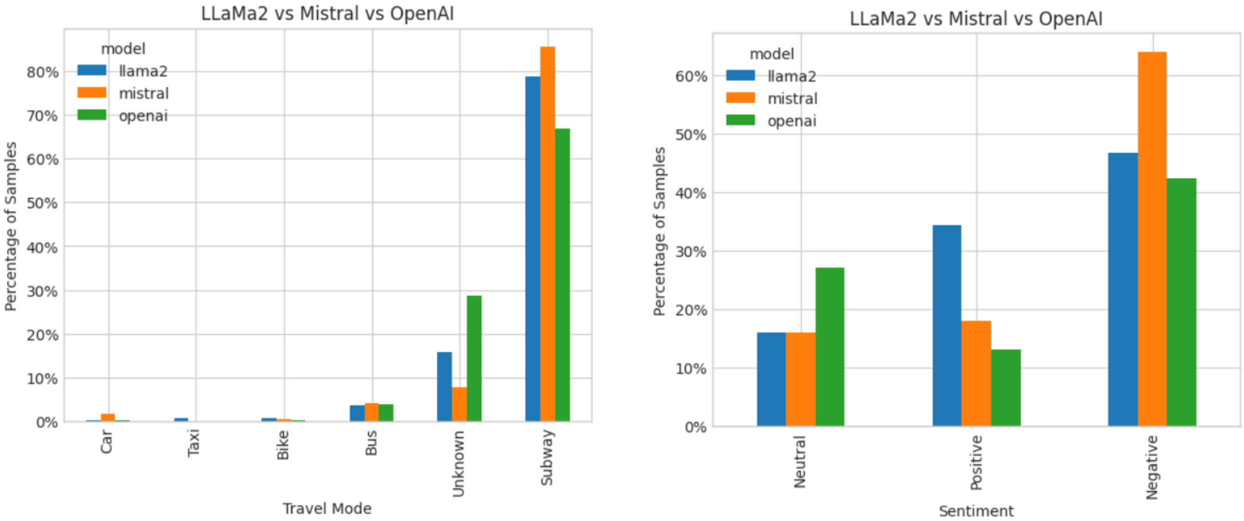
Figure 3: Llama2 vs Mistral on all 2000 samples

Finally, the reasonings provided by different models were compared using cosine similarity scores and plotted against the index. Figure 5 shows cosine similarity scores between the LLama2 and OpenAI's reasonings. Observe that the scores are approximately bell-shaped and have little to no scoes above 0.5 in magnitude. This shows that the models (Llama2 and GPT-3.5-Turbo) have very different architectures and peculiar ways of reasoning for a given statement.

More detailed comparisons can be found in the *comps* notebook.

7 Conclusion

In summary, we were able to generate predictions on all the 2000 samples using 2 models. Verification with OpenAI as well as manual human evaluation was conducted. The predictions made by



(a) Mode of Travel

(b) Tweet Sentiment

Figure 4: Llama2 vs Mistral vs OpenAI on all 1200 samples

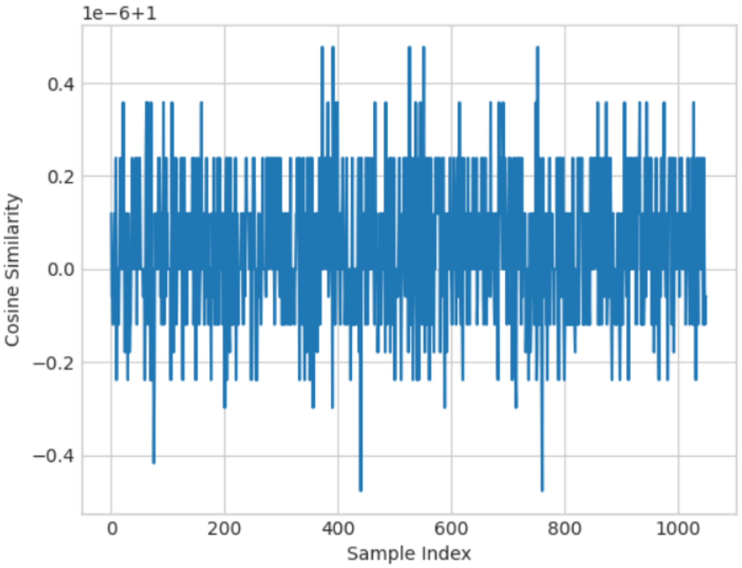


Figure 5: Reasoning Alignment between Llama2 and OpenAI's GPT-3.5-Turbo

Llama2 and Mistral agree 84% of the times. This builds trust in the algorithm of detecting travel mode and the corresponding sentiment. I propose that we merge the model predictions for the pair *Unknow and Subway*, made by llama2 and mistral to get the final mode of travel. Since, this change was observed during the human evaluation step, it can be proved to remove false negatives and boost the overall accuracy.

8 Acknowledgments

I would like to thank Darren Ruan for his support and guidance throughout the semester. I would also like to thank him for providing me with GCP credits to experiment with various LLMs on efficient GPUs.

References

[1] X. Chen, Z. Wang, and X. Di, "Sentiment analysis on multimodal transportation during the covid-19 using social media data," *Information*, vol. 14, no. 2, p. 113, 2023.

[2] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer, "Rethinking the role of demonstrations: What makes in-context learning work?" *arXiv preprint arXiv:2202.12837*, 2022.

[3] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.