TransportGPT

Prompt Augmented Generation

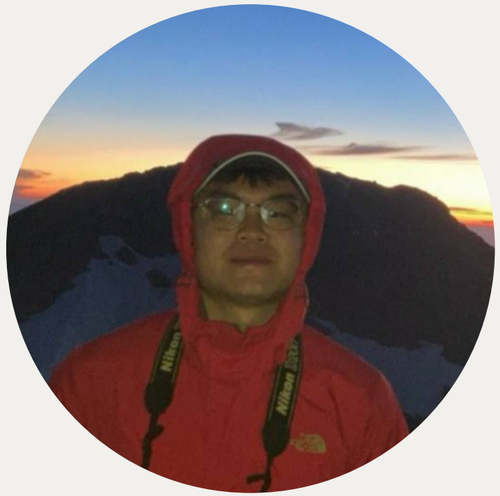
Team TAB

Meet TAB

Elias

Koko

Jden



Shu Heng

Javier

Prompt Augmented Generation

Retrieval Augmented Generation (RAG) [1] is commonly used to embed context to prevent LLM hallucinations. To combat constraints in term of lack of compute and proper labelled data, we propose **Prompt Augmented Generation** instead.

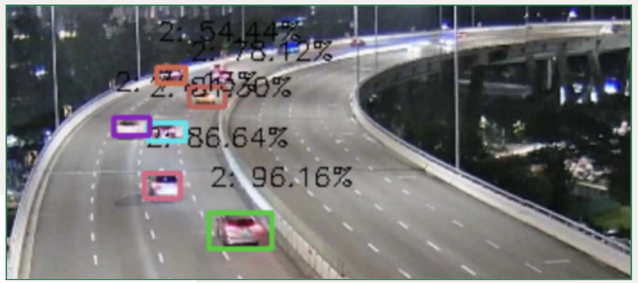
Instead of interacting with the input embeddings, we use the ensemble of data sources to augment the prompt instead.

[1] P. Lewis, E. Perez, A. Piktus, Fabio Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", Conference on Neural Information Processing Systems (NeuRIPS), 2020

Prompt Augmentation Process

$$\mathcal{D}_{rel} = \underset{i \in |\mathcal{D}|}{\operatorname{argmin}} \sqrt{\|\mathcal{D}_i - \mathcal{D}_{user}\|_2^2}$$

We conduct **1-stage Information Retrieval on live data** from LTA DataMall via **KNN-search** on the vector representation of the driver's starting and destination coordinates.



We use LTA API to access live traffic camera feeds and conduct **object detection via YoloV3** to determine the road conditions near the driver.

Camera Feeds

Land Transport DATAMALL

Numerical Data

Pathos

Front End

NEXT.JS

Google Maps

FastAPI

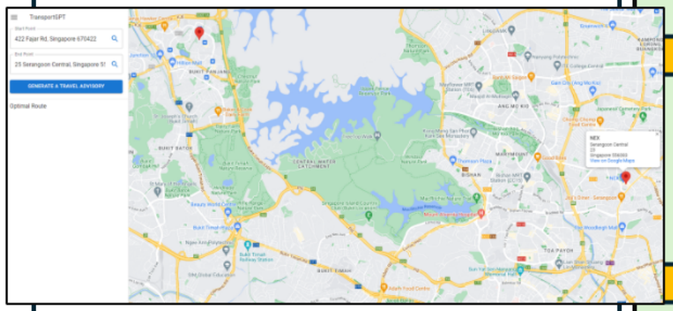
GeoPy

scikit learn

OpenCV

Retrieval Engine

The retrieval engine pulls data from the camera feeds and LTA based on the user's input.



LangChain

Augmentation

Inference

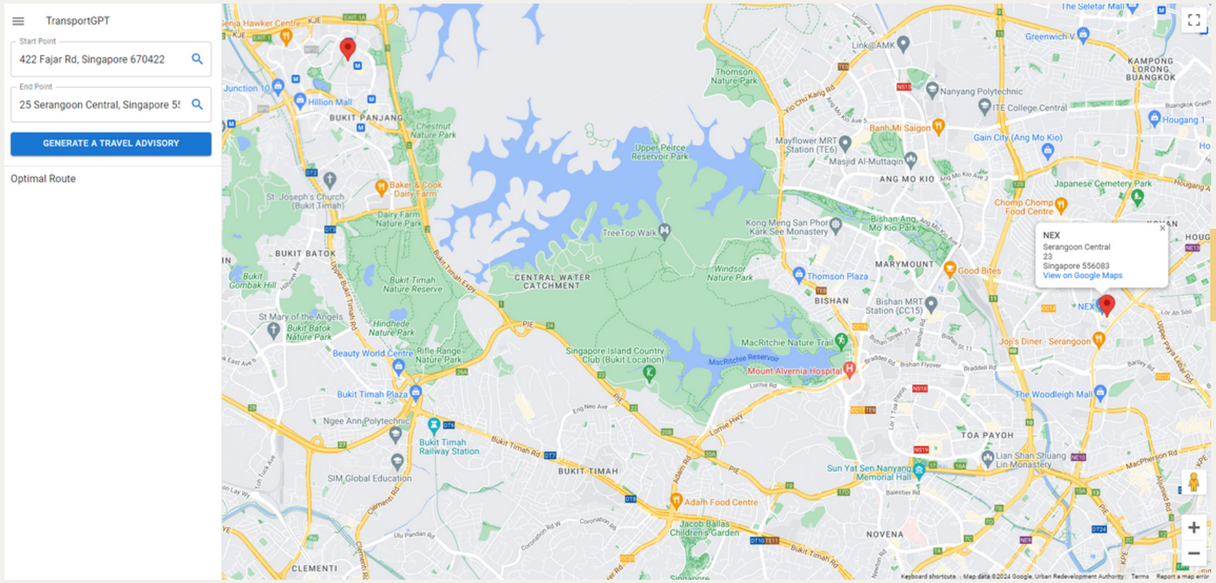
OpenAI

Our web-app integrates directly into Google Maps

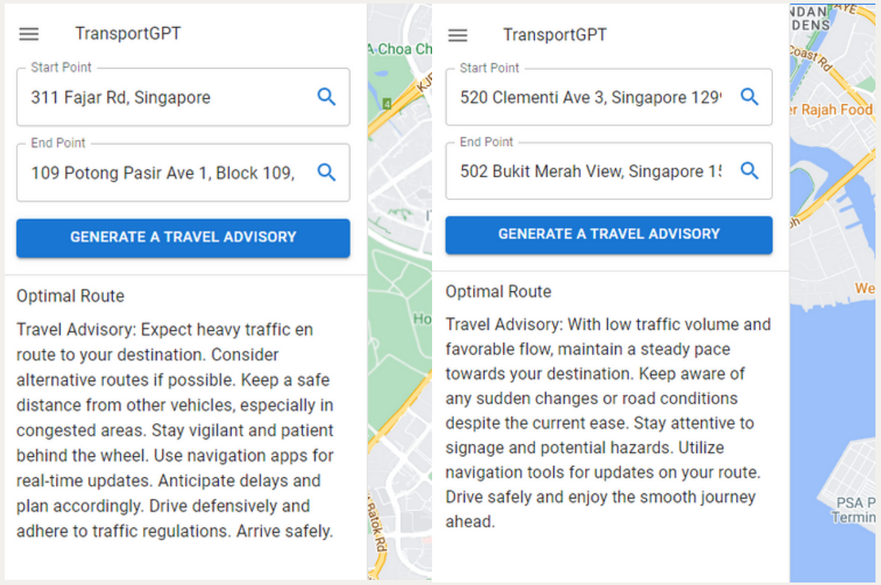
The various data sources are used to augment the user's original prompt via LangChain.

User Experience

To **enhance user experience & ease up adoption**, we seamlessly integrated our application into a user-friendly platform that most drivers are already familiar and comfortable with: **Google Maps**.



TransportGPT working alongside GoogleMaps
Uses Reverse Geocoding to obtain
the start and end locations that the user selected



We can see TransportGPT taking traffic conditions into consideration during inference.

Benefits

- **Real-Time Updates:** Drivers receive real-time, context-aware traffic updates
- **Improved Navigation:** The app provides navigation advisories based on real-time road conditions
- **Enhanced Safety:** Alerts drivers to potential hazards in real-time contributing to safer driving conditions and reduces the risk of accidents.

Further Work

We plan to take more time to create labelled dataset, together with LoRA [1], we would be able to explore this platform in a more traditional RAG architecture, possibly exploring a 2-stage **re-ranking** retrieval process to improve performance**.**

We also plan to explore a **Mixture-Of-Experts (MoE)** for the retrieval / augmentation step to **support the users in more downstream tasks** apart from generating traffic advisories.

The platform could be expanded to support:

Continuous Learning: We can attempt this via LoRA arithmetic [2] to ensure **scalable and sustainable** online learning whilst **handling catastrophic forgetting.**

Mixture-Of-LoRA [3]: LoRA also allows for us to conduct **low-compute** Mixture-Of-Experts.

[1] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models", 10th International Conference on Learning Representations (ICLR), 2022
[2] R. Chitale, A. Vaidya, A. Kane, A.S. Ghotkar, "Task Arithmetic with LoRA for Continual Learning", Conference on Neural Information Processing Systems (NeuRIPS), 2023
[3] W. Feng, C. Hao, Y. Zhang, Y. Han, H. Wang, "Mixture-of-LoRAs: An Efficient Multitask Tuning for Large Language Models", International Conference on Computational Linguistics (COLING), 2024

Thank you