TransportGPT: Prompt Augmentation Generation

Elias Lim, Moreno Koko, Javier Ng, Jden Goh, Yeo Shu Heng

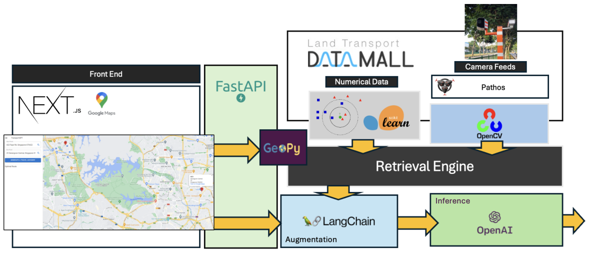


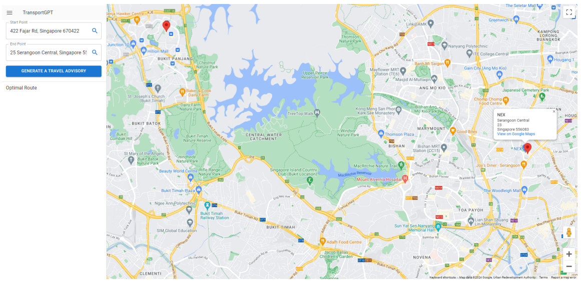
Fig. 2. Proposed Architecture for Prompt Augmentation Generation

I. Introduction

Large Language Models (LLM) show SOTA performance on NLP tasks, however, they fail when task/context-specific requirements are present. Retrieval Augmented Generation (RAG) [1] allows LLM to work on a given task and prevent hallucinations by augmenting the input embedding with additional ground truth data the model should consider.

II. Architecture

Due to lack of time to create good datasets, we abandoned a traditional RAG-like framework, but carry over the intuition by injecting context through a *augmented prompt* over an ensemble of downstream data. Augmentation is done via LangChain. We integrated it with Google Maps to give drivers a platform they are familiar and comfortable with for seamless adoption.



TransportGPT

Start Point: 422 Fajar Rd, Singapore 670422

End Point: 25 Serangoon Central, Singapore 5!

GENERATE A TRAVEL ADVISORY

Optimal Route

Fig. 1. TransportGPT works seamlessly with Google Maps

III. Performance & Results

A. Ensemble of Data

Continuous data from LTA data mall follows a 1-stage retrieval on journey-relevant features (Estimated Travel Time, Traffic Flow, Roadworks) by transposing the user's position and destination into a coordinate vector $\mathcal{D}_{user}$ and using KNN against the coordinate vector representation of LTA data mall points to obtain the most relevant datapoint $\mathcal{D}_{rel}$ via

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

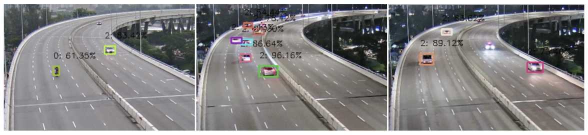
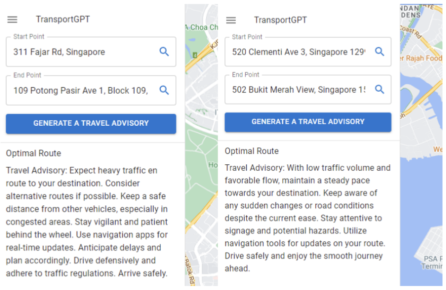


Fig. 3. YoloV3 shows consistent performance across lighting conditions.

Live camera feeds, alongside YoloV3, supplements the augmentation process by providing the density of vehicles within user's vicinity. The image processing pipeline could be extended to consider of road works, type of vehicles and road size in the future. We tried to ensure real-time inference via multiprocessing.

B. Prompt Augmentation

The following displays some prompts generated from our augmentation process, we can see traffic conditions being considered in the output.



TransportGPT

Start Point: 311 Fajar Rd, Singapore

End Point: 109 Potong Pasir Ave 1, Block 109,

GENERATE A TRAVEL ADVISORY

Optimal Route

Travel Advisory: Expect heavy traffic en route to your destination. Consider alternative routes if possible. Keep a safe distance from other vehicles, especially in congested areas. Stay vigilant and patient behind the wheel. Use navigation apps for real-time updates. Anticipate delays and plan accordingly. Drive defensively and adhere to traffic regulations. Arrive safely.

TransportGPT

Start Point: 520 Clementi Ave 3, Singapore 129

End Point: 502 Bukit Merah View, Singapore 1!

GENERATE A TRAVEL ADVISORY

Optimal Route

Travel Advisory: With low traffic volume and favorable flow, maintain a steady pace towards your destination. Keep aware of any sudden changes or road conditions despite the current ease. Stay attentive to signage and potential hazards. Utilize navigation tools for updates on your route. Drive safely and enjoy the smooth journey ahead.

Fig. 4. Outputs generated by TransportGPT across various road conditions.

IV. Further Work

With more time for data collection, we hope to explore a more RAG-like structure, with augmentation conducted on input-embedding. We could add a separate retrieval engine and train in tandem with PEFT on the LLM via LoRA (Low Rank Adaptation) [2] to cope with compute. Generalization gaps due to LoRA could be avoided by improving its learning capacity using various techniques [3]. LoRA also allow flexibility to extend into low-compute incremental learning deployments [4] and increase task flexibilty via a MoA [5] architecture, allowing our platform to be scalable.

V. Bilbography

The following consists of sources & references used in the construction of our report.

References

[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

[2] 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

[3] B. Zi, X. Qi, L. Wang, J. Wang, K. Wong, L. Zhang, ”Delta-LoRA: Fine-Tuning High-Rank Parameters with the Delta of Low-Rank Matrices”, *arXiv:2309.02411 [cs.LG]*, 2023

[4] R. Chitale, A. Vaidya, A. Kane, A.S. Ghotkar, ”Task Arithmetic with LoRA for Continual Learning”, *Conference on Neural Information Processing Systems (NeuRIPS)*, 2023

[5] 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