Jing YANG, Xingyuan DAI, Yisheng LV, Levente KOVÁCS, Fei-Yue WANG, 2025. TransRAG for parallel transportation: toward reliable and trustworthy transportation systems via retrieval-augmented generation. *Frontiers of Information Technology & Electronic Engineering*, 26(1):20-26. https://doi.org/10.1631/FITEE.2400800

TransRAG for parallel transportation: toward reliable and trustworthy transportation systems via retrieval-augmented generation

Key words: Parallel transportation; Foundational models; Blockchain; Caching; Smart contracts; Retrieval-augmented generation

Corresponding author: Fei-Yue WANG
E-mail: feiyue.wang@ia.ac.cn
ORCID: http://orcid.org/0000-0001-9185-3989

Motivation

Recently, the advancement of foundation models has elevated the implementation of parallel transportation systems to a new level because of their exceptional understanding, generation, and reasoning abilities. However, the inherent flaws of foundational models (FMs) hinder their application in actual transportation systems and could even jeopardize the safety of passengers and drivers as follows:

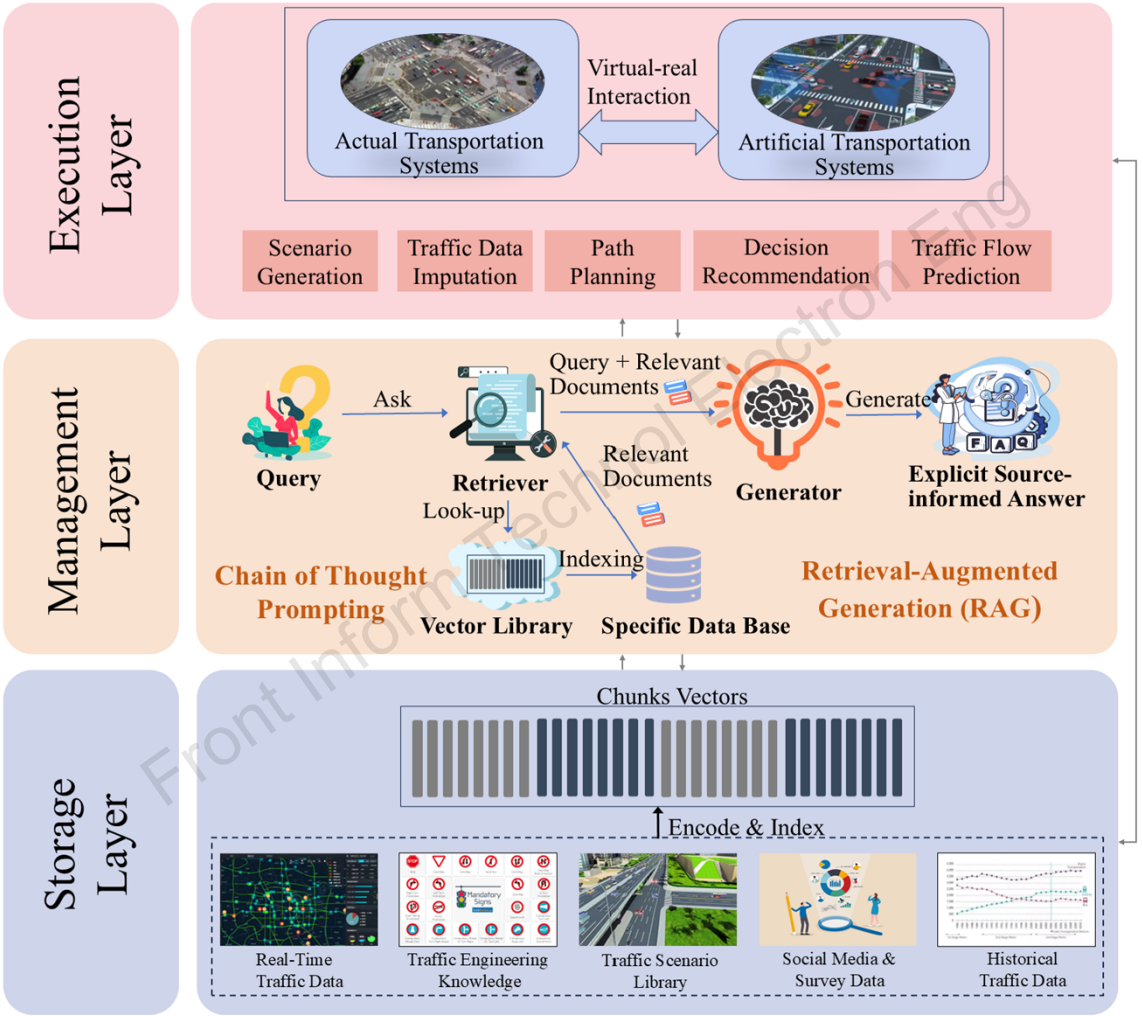
(1) “Hallucination” phenomena and outdated knowledge within FMs contribute to the unreliability of their inferences.

(2) The fact that FMs are neural network based “black-box” models makes their generation results inexplicable and untrustworthy.

Main idea

- We propose a unified TransRAG framework for parallel transportation to guarantee the dependability and trustworthiness of FMs' decision-making in different traffic scenarios and thus the safety of traffic system operations.
- Retrieval-augmented generation (RAG) and chain-of-thought (CoT) prompts are integrated into the framework to enable FMs to access real-time knowledge and information without the need for re-pretraining, steering them toward generating more accurate and interpretable outcomes.

Framework



Framework of TranRAG (RAG: retrieval-augmented generation)

Storage layer

- The storage layer is the lowest tier, primarily responsible for storing actual data and knowledge and supporting web-based retrieval functions.
- The stored data can be updated in real time, including real-time traffic data, traffic engineering knowledge, traffic scenario library, historical traffic data, and social media & survey data.
- To speed up retrieval and meet FMs' input token limits, all the texts in the database are broken down into several smaller and more approachable chunks. Subsequently, an embedding model is used to encode these chunks as corresponding vector representations, and thus an index is created where chunks are stored as values and vector representations as keys.

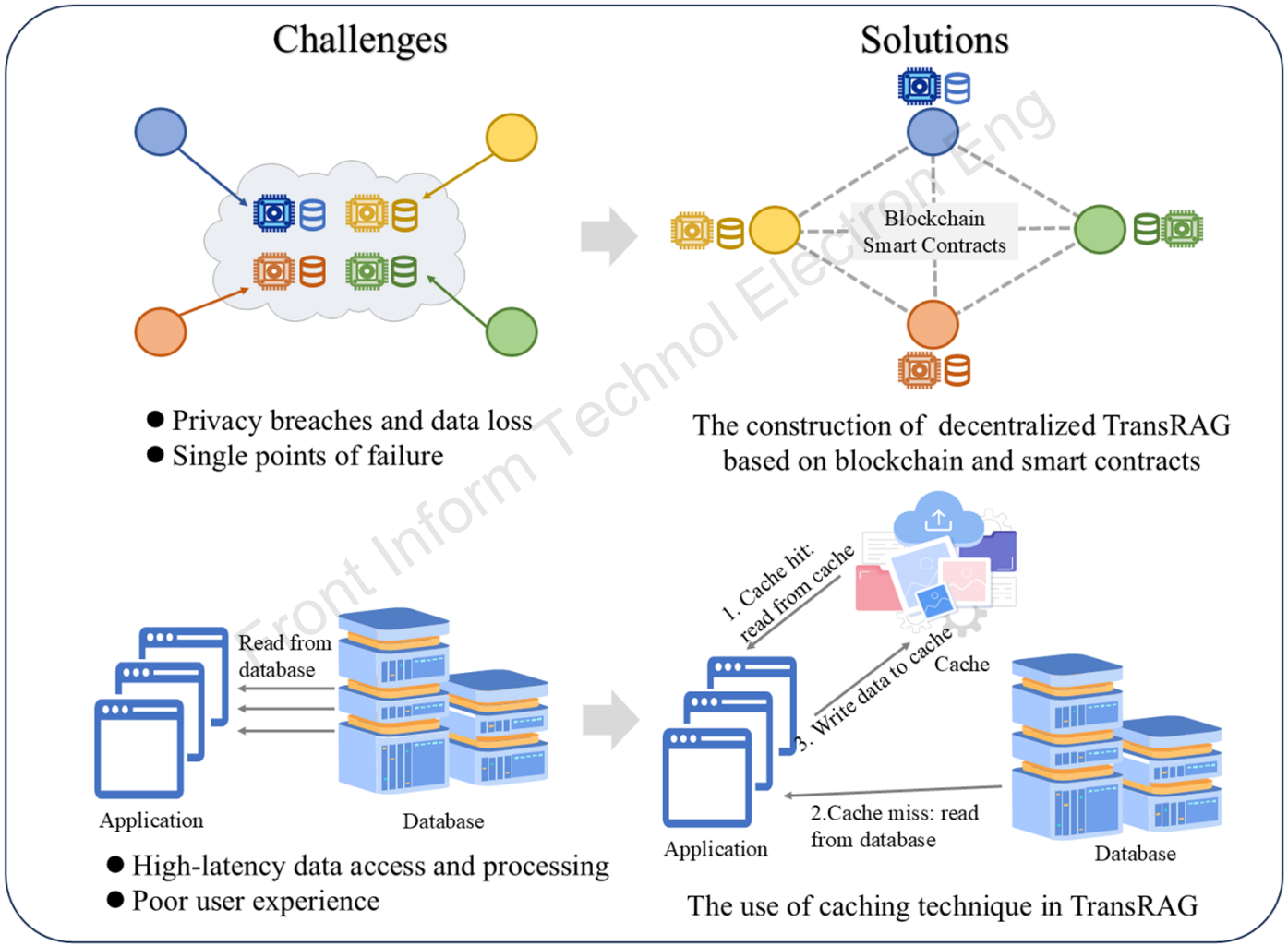
Management layer

- The management layer is the middle tier and leverages RAG and CoT to guide foundation models in making reliable and interpretable decisions through computational experiments.
- The entire decision-making process can be divided into two stages, retrieval and generation, so the layer includes at least one retriever and one generator.
- In the retrieval stage, after a query is asked, the same embedding model as the storage layer is leveraged to encode it as a query vector. The top-*K* most similar chunk vectors are identified and returned as relevant information by computing the similarity between the query vector and the chunk vectors.
- In the generation stage, the query, its relevant information, and task-specific prompts are combined as a coherent contextual prompt, fed into the generator to formulate a response.

Execution layer

- The execution layer is the highest tier, where the decisions from the management layer are executed in actual transportation systems in parallel with artificial transportation systems.
- A variety of artificial systems are built using scenario data from the storage layer, where computational experiments are performed by the management layer.
- During parallel execution, the differences between the two types of systems are fed back into the management layer to continuously optimize decision-making.
- Generated intermediate data and tested case studies are transmitted to the storage layer for storage, thereby accumulating experience for future strategy formulation.

Opportunities and challenges



Conclusions

- Transportation 4.0 to Transportation 5.0 requires a unified approach that seamlessly integrates all these elements, especially human and societal factors, instead of addressing them as separate issues.
- The proposed TransRAG for parallel transportation shows promise in advancing the shift toward Transportation 5.0 as a holistic framework.
- The shortcomings of TransRAG, such as privacy breaches, single point of failure, and delays in data access, are analyzed, and the corresponding solutions are offered, including the use of blockchain, smart contracts, and caching.

Jing YANG received the bachelor degree in automation from Beijing University of Chemical Technology in 2020. She is currently a PhD candidate at the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China. Her research interests include parallel manufacturing, social manufacturing, cyber–physical–social systems, and artificial intelligence.

Xingyuan DAI is an associate professor at State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, China. He received the PhD degree in pattern recognition and intelligent system from Institute of Automation, Chinese Academy of Sciences in 2012. He was a visiting scholar at University of California, Berkeley in 2019. His research interests include data mining and machine learning.

Yisheng LV received the BE and ME degrees in transportation engineering from Harbin Institute of Technology, Harbin, China, in 2005 and 2007, respectively, and the PhD degree in control theory and control engineering from the Chinese Academy of Sciences, Beijing, China, in 2010. He is currently a Professor with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. His research interests include traffic data analysis, dynamic traffic modeling, and parallel traffic management and control systems.

Levente KOVÁCS received the PhD degree in electrical engineering from the Budapest University of Technology and Economics, Budapest, Hungary, in 2008. He is Full Professor with the John von Neumann Faculty of Informatics, Óbuda University, Budapest. He is the Rector/President with Óbuda University. He has authored and coauthored more than 500 articles in international journals and refereed international conferences, with an *H*-index of 30. He founded the Physiological Controls Research Center at Obuda University in 2013, and was its Head. He was János Bolyai Research Fellow of the Hungarian Academy of Sciences during 2012–2015. His research interests include modern control theory and physiological controls.

Fei-Yue WANG received the PhD degree in computer and systems engineering from Rensselaer Polytechnic Institute, Troy, NY, USA, in 1990. He is currently a Professor with the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China, also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, and also with the Macau Institute of Systems Engineering, Macau University of Science and Technology, Macau, China. He received the IEEE ITS Outstanding Application and Research Awards in 2009, 2011, and 2015, and the IEEE SMC Norbert Wiener Award in 2014. In 2021, he became the IFAC Pavel J. Nowacki Distinguished Lecturer. He is a Fellow of the International Council on Systems Engineering, International Federation of Automatic Control, American Society of Mechanical Engineers, and American Association for the Advancement of Science. His current research focuses on methods and applications for parallel systems, social computing, and knowledge automation.