

Enhancement of Large Language Models Driving Knowledge for Practical Autonomous Driving Decision Making

Dashuai Pei, Jianhua He, *Senior Member, IEEE*, Kezhong Liu, Yiwen Wu, Yaxiong Lei, Xuedou Xiao

Abstract—Theoretical driving knowledge and hazard perception capability are essential for large language models (LLMs) to be qualified for driving or assisting autonomous vehicles. Our prior study found that the mainstream LLMs with relatively small size suitable for onboard or edge deployment failed driving theory tests. That raised several important research questions on application of LLMs for safety critical autonomous driving tasks: how to enhance the theoretical driving knowledge of LLMs and how to assess its impact on driving decision-making and safety in realistic scenarios. In this paper we first propose a novel question guided multimodal graph based retrieval-augmented generation (QGM-GRAG) approach to address the first question. Structured textual rules and visual traffic sign semantics are integrated to create a unified graph-based representation. A joint representation that fuses paragraph level embeddings with question aware signals is generated by LLMs for accurate and interpretable retrieval. Then we propose a general scenario based evaluation framework to address the second question, by investigating the connection between the driving theory knowledge and practical driving decision-making performance. In the framework theory-test questions are enhanced and converted to realistic driving scenarios. Prompts are designed for LLMs to generate driving decisions. The decisions are evaluated automatically by a GPT-4o based agent across multiple dimensions (such as situation awareness and driving safety). Experiments demonstrate that the proposed RAG approach can achieve more than 10% improvement on driving theory tests. In addition, results from the evaluation framework show that the driving knowledge has significant impact on driving decisions and the QGM-GRAG approach can help improve driving performance by more than 13%. Furthermore, experiments on the real-world LingoQA dataset show an average improvement of 44.87%, confirming the practical applicability of our pipeline. These results validate the effectiveness and robustness of the proposed RAG approach and evaluation framework.

Index Terms—LLMs, Multimodal RAG, Driving Theory Tests, Theory-to-behavior Transfer.

I. INTRODUCTION

Large Language Models (LLMs), such as GPT-4 [1] and its contemporaries, have attracted increasing attention

for autonomous driving applications due to their unprecedented capabilities of statistical inference-like processing, language understanding, and decision-making. Recent studies have demonstrated the potential of LLMs to assist or even lead critical tasks in autonomous vehicles, particularly in complex scenarios which require common sense inference, multimodal understanding, and human-like decision generation. The applications of LLMs to autonomous driving can be grouped to two broad tasks: LLM for driving decision-making and LLMs for driving assistance [2]–[4]. The former explores using LLMs as core decision inference agents capable of interpreting sensor inputs and generating high-level driving actions or tactical decisions, and the latter focuses on leveraging LLMs to enhance driving safety and compliance by providing explainable decision-making, behavior justification, and human-in-the-loop assistance. These applications could be deployed both onboard or remotely (in the cloud or at the network edge). As the cloud deployment could introduce unacceptable network delay and communication traffic, LLMs deployment on vehicles or at the edge is more practical and feasible. That favors the use of small-footprint LLMs (such as with 1B to 10B parameters) to meet real-world constraints on computation, latency, energy consumption, and hardware compatibility [5], [6].

Despite the potentials of LLMs on autonomous driving and a large body of work investigating of its feasibility with data-driving training, an important aspect that has been largely underexplored is the assessment of the driving domain knowledge of the LLMs and exploitation of the driving theory knowledge for practical driving decision-making. For the safety-critical autonomous driving applications, we argue that the LLMs should possess strong theoretical driving knowledge and hazard perception capability before they can be qualified for driving or assisting autonomous vehicles. Such requirements should be similar to or even stricter than those for human drivers, who need to pass mandatory driving theory tests and practical road tests to obtain driving licenses. Our previous work on driving theory tests of mainstream LLMs demonstrated that small-scale models (with fewer than 10 billion parameters), which are otherwise attractive for onboard or edge deployment, failed to achieve acceptable accuracy on the UK driving theory benchmark [7]–[9]. It indicates significant gaps in their driving domain knowledge and rule-based decision inference abilities.

The observation from our previous study raises several critical questions that should be addressed to successfully

Dashuai Pei (email: Pei.Dashuai@whut.edu.cn) and Xuedou Xiao (xuedouxiao@whut.edu.cn) are with the school of navigation, Wuhan University of Technology (WUT), Wuhan, China. Jianhua He (email: j.he@essex.ac.uk) is with the school of computer science and electronic engineering, University of Essex, UK. Kezhong Liu (email: kzliu@whut.edu.cn) is with the school of navigation and state key laboratory of maritime technology and safety, WUT, Wuhan, China. Yiwen Wu (email: YiwenWu@ieee.org) is with the school of computer science, University of Electronic Science and Technology of China. Yaxiong Lei (yl212@st-andrews.ac.uk) is with University of St Andrews and University of Essex, UK.

Jianhua He and Kezhong Liu are corresponding authors.

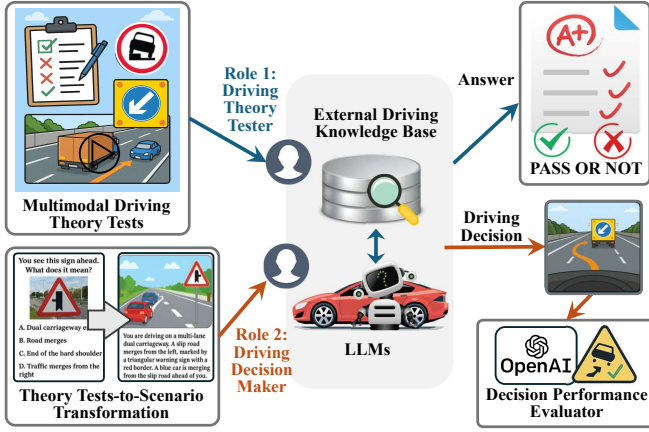


Fig. 1. Overview of our work. We first enhance small LLMs with a multimodal driving knowledge base via a targeted RAG pipeline to improve their theoretical test performance. Subsequently, we transform theory questions into realistic driving scenarios and evaluate the resulting decisions through GPT-4o-based multi-dimensional scoring.

apply and integrate LLMs into autonomous driving systems. First, how can the performance of LLMs on driving theory tests be effectively improved with enhanced driving knowledge, for small-scale models that lack sufficient domain-specific knowledge? Given that retraining or fine-tuning LLMs on extensive driving-specific corpora can be very expensive and computationally intensive, more efficient approaches are needed to inject essential driving knowledge without compromising deployment feasibility [10], [11]. Second, how can we exploit the driving domain knowledge for practical driving and assess its impact on driving decision-making performance (such as safety and efficiency)? Rather than training LLMs with large corpora of sensor and driving data based samples, training LLMs with structured driving knowledge data samples is expected to be much more efficient. However, there is a lack of work investigating such approach. Third, how much improvements in theory test performance can be translated into driving decision-making performance in realistic and complex scenarios? While a higher theory test score may reflect better factual knowledge, it remains unclear whether these gains will be reflected in practical driving situations, where context-sensitive, legally compliant, and risk-aware decisions should be generated under uncertainty. Addressing these questions is essential to bridge the gap between theoretical competence (knowledge of traffic laws and safe practices) and operational competence (contextual decision-making and behavior generation), ultimately supporting efficient training and application of LLM-based agents for safety-critical driving applications.

In this paper we investigate the LLMs driving knowledge enhancement and its impact on practical driving, which is illustrated in Fig. 1. We first design a new question-guided multimodal graph retrieval-augmented generation (QGM-GRAG) approach to address the first question. Compared to costly full-model pretraining or fine-tuning, RAG offers a lightweight and scalable alternative, especially suitable for small-scale models deployed on vehicles. By retrieving relevant driving rules, procedures, or best-practice demonstrations from a curated

knowledge base, RAG-equipped models can compensate for their limited internal capacity and improve factual correctness and regulatory compliance [12]. However, directly applying conventional RAG pipelines presents major challenges. As driving theory questions are often short, interrogative prompts with limited lexical cue will reduce retrieval precision. In addition, many questions include images (e.g., road signs or junction layouts), the standard text-only RAG systems cannot be used directly to work with these driving questions. Furthermore, even if theory-level reasoning is improved, exploiting the static knowledge from RAG for dynamic and context-sensitive driving decision-making remains very challenging. Our proposed QGM-GRAG approach can address the limitations of the existing RAG methods. Structured textual rules and visual traffic sign semantics are integrated to create a unified graph-based representation. A question-guided joint embedding is generated by LLMs to improve semantic retrieval performance across modalities. Then we design a general scenario-based assessment framework to address the second and third questions. Theory-test questions with improved answers are transformed into realistic, context-rich driving scenarios, where LLMs are prompted to generate safe and compliant actions. A GPT-4o-based evaluation module then scores each decision across multiple dimensions (such as legal compliance, safety awareness, and behavioral realism). It can be used to exploit the driving domain knowledge for practical road driving and evaluate the theory-to-behavior transfer. The framework is applicable to general LLMs and methods that enhance LLMs' theoretical understanding. It can also be used to assess whether those improvements translate into better real-world decision-making.

Our contributions are summarized as follows:

- 1) We formulate the problem of systematically enhancing LLMs' driving-theory competence and evaluating its impact on decision-making. To address this problem, we propose a question-guided multimodal graph-based retrieval framework (QGM-GRAG) that integrates structured regulatory text and traffic-sign semantics into a unified knowledge graph representation. By introducing a joint embedding mechanism that combines paragraph-level semantic representations with LLM-generated question signals, the framework improves alignment between natural-language queries and grounded driving knowledge.
- 2) We develop a general scenario-based evaluation framework that transforms theory-level knowledge into structured driving scenarios and assesses decision quality along multiple interpretable dimensions. Unlike conventional evaluation protocols that focus solely on answer correctness, this framework explicitly measures whether improvements in theoretical understanding translate into measurable gains in decision-level behavior.
- 3) Through controlled experiments across multiple LLM architectures and retrieval configurations, we provide a systematic empirical analysis linking driving-theory test performance with scenario-level decision behavior. The results reveal a consistent relationship between enhanced

factual and procedural knowledge and improved decision accuracy.

Together, these contributions offer a unified pathway for strengthening and evaluating LLMs-based driving intelligence, from knowledge acquisition to operational decision-making, under real-world constraints, marking a step forward in the integration of LLMs into safe, explainable autonomous driving systems. By enhancing the theoretical and behavioral competence of LLMs for vehicle-level decision intelligence, this work supports the advancement of safer and more interpretable components within broader intelligent transportation systems involving multiple vehicles, infrastructure, and road users.

The remainder of this paper is structured as follows. Section II reviews existing research on applying LLMs to autonomous driving tasks. Section III presents the proposed QGM-GRAG framework in detail. In Section IV, we introduce our scenario-based evaluation framework. Comprehensive experimental results validating the effectiveness of our retrieval-augmented generation approach and evaluation methodology are reported in Section V. Finally, Section VI summarizes our key findings.

II. RELATED WORK

A. LLMs in Driving Theory and Decision-Making

LLMs are increasingly studied for driving applications, both as core decision-making agents and as auxiliary inference components. Prior studies primarily fall into two categories: (1) using LLMs for driving decision-making, and (2) using LLMs for driving assistance and theory comprehension.

In the first paradigm, researchers explore how LLMs, particularly vision-language models, can interpret multimodal sensor data (e.g., camera feeds, radar maps) and generate high-level driving intentions. These systems often involve LLMs producing natural language explanations, situational summaries, or meta-decisions that guide downstream planning modules. Notable early works such as DriveGPT [13], OpendriveVLA [14], and LMDrive [15] have explored how LLMs can interpret visual inputs and generate natural language driving commands, often by pairing camera feeds with human-written instructions or demonstration labels. These approaches typically frame driving as a sequence prediction or language generation task, grounded in perception. While promising in low-speed or simulation settings, these models often lack grounding in explicit driving rules and safety constraints, which limits their trustworthiness in real-world deployment.

In the second paradigm, LLMs are applied to driving assistance tasks, such as answering UK DVSA-style driving theory questions. These evaluations provide insight into a model’s grasp of traffic law, road signs, safety procedures, and ethical driving behavior. Furthermore, few studies have attempted to connect these two paradigms by testing whether better theory comprehension (e.g., knowing when to yield) leads to better decision-making (e.g., stopping at a junction). Our work fills this void by constructing a structured evaluation framework that explicitly measures this theory-to-behavior transfer and investigates how knowledge integration strategies like RAG can improve not only factual performance but also real-world safety reasoning.

B. Knowledge Graphs for Driving Decision-Making

Knowledge graphs (KGs) have been widely investigated as a way to make driving decision pipelines more structured, queryable, and rule-compliant, especially when perception outputs must be grounded in explicit traffic semantics and regulations. For instance, CoSI [16] proposed a KG-based “coherence layer” that fuses heterogeneous driving-context factors and then exploits axioms and reasoning to derive semantically enriched situation understanding signals. The work demonstrates how an explicit ontology + reasoning stack can complement learning-based modules by providing structured context that is easier to interrogate and extend to new knowledge sources. More recently, graph-structured knowledge has been used to scaffold multimodal reasoning for driving decisions. DriveLM [17] introduced Graph Visual Question Answering, where perception–prediction–planning questions are linked via logical dependencies over an implicit scene graph, enabling stepwise reasoning that better matches how human drivers justify actions. This line of work positions “graph-structured” intermediate representations as a bridge between raw perception and decision outputs, improving generalization and interpretability. Finally, KG-based retrieval has been explored to inject domain safety knowledge into vision-language driving assistants. SafeDriveRAG [18] builds a multimodal traffic-safety knowledge graph from documents and retrieves relevant subgraphs to support safety-critical question answering and reasoning, reporting consistent gains over base VLMs on a large-scale safety QA benchmark. This supports the broader view that KGs are increasingly used not only for static world modeling, but also as an actionable memory structure for retrieval-augmented, constraint-aware driving decisions.

While these studies demonstrate the value of KGs for structured reasoning, scene understanding, or safety-oriented retrieval, they primarily focus on perception-grounded scene graphs, ontology-driven symbolic reasoning, or safety QA over traffic documents. In contrast, our work addresses a different and largely underexplored problem: how to systematically enhance large language models’ driving-theory competence and explicitly evaluate its downstream impact on driving decision-making. Rather than constructing a scene graph from perception outputs or relying on predefined ontological reasoning rules, we build a question-guided multimodal KG directly from structured driving regulations and traffic-sign semantics, and integrate it into a RAG framework. Moreover, we propose a scenario-based transformation mechanism that connects theory-level knowledge to practical decision contexts, enabling us to quantify how improvements in theoretical grounding translate into measurable gains in decision quality. This explicit linkage between theory enhancement and decision performance, supported by a unified multimodal KG and evaluation pipeline, distinguishes our framework from prior KG-based driving systems.

C. RAG for Autonomous Driving

RAG has emerged as a practical approach to supplement the factual knowledge of LLMs at inference time, without requiring additional fine-tuning [19]. Initially developed for

open-domain question answering, RAG methods have since been extended to a variety of domain-specific tasks where models are expected to reason with structured or technical knowledge not fully covered in pretraining corpora. Applications include biomedical QA [20], legal document summarization [21], and technical instruction following [22]. In the driving domain, RAG techniques have recently gained attention for their ability to integrate external, structured information into decision-making pipelines. For instance, Driving with Regulation [12] introduced a retrieval-augmented reasoning framework that incorporates traffic regulations into autonomous vehicle planning. Their system retrieves relevant legal clauses and uses LLM-based reasoning to generate interpretable, regulation-compliant driving decisions. The work highlights the importance of aligning LLM decision outputs with domain-specific safety and legal constraints, particularly in high-stakes environments like road navigation. Similarly, SenseRAG [23] used sensor-derived observations to retrieve context-aware knowledge, improving hazard anticipation and intention prediction. These works illustrate the potential of RAG to enhance LLMs' performance in sensor-grounded or multimodal driving systems.

However, existing work, including Driving with Regulation [12], primarily emphasizes in-the-loop planning and interpretable control, often assuming a well-defined scene and retrievable legal support. By contrast, our work focuses on the knowledge acquisition gap itself: we ask whether retrieval-augmented LLMs can first learn to internalize structured traffic knowledge, and then apply it in new, unstructured driving scenarios. We evaluate whether improvements on structured driving theory tests translate into improved downstream decision-making. This cognitive transfer, from static regulation understanding to dynamic situational reasoning, has not been explicitly measured in prior RAG-based autonomous driving research.

III. QGM-GRAG APPROACH

A. Overview of QGM-GRAG

We propose QGM-GRAG, a novel approach for question-guided, cross-modal RAG, as shown in Fig. 2. The approach consists of two main components: (1) multimodal knowledge graph construction, and (2) question-guided multimodal graph retrieval. The core idea is to integrate textual and visual driving knowledge within a unified semantic graph structure, enabling effective question-guided retrieval.

Specifically, the resulting knowledge graph consists of three primary node types. First, text nodes represent paragraph-level regulatory content derived from the driving rule corpus and are enriched with question-guided joint embeddings that combine paragraph semantics with generated question intent. Second, image nodes correspond to traffic-sign images encoded in the CLIP embedding space. Third, description (SBN) nodes represent textual traffic-sign descriptions embedded in the same E5 semantic space as text nodes, serving as semantic bridging anchors across modalities. Edges are defined according to this structured schema. Text-text edges are weighted by a combination of semantic similarity and concept overlap to capture

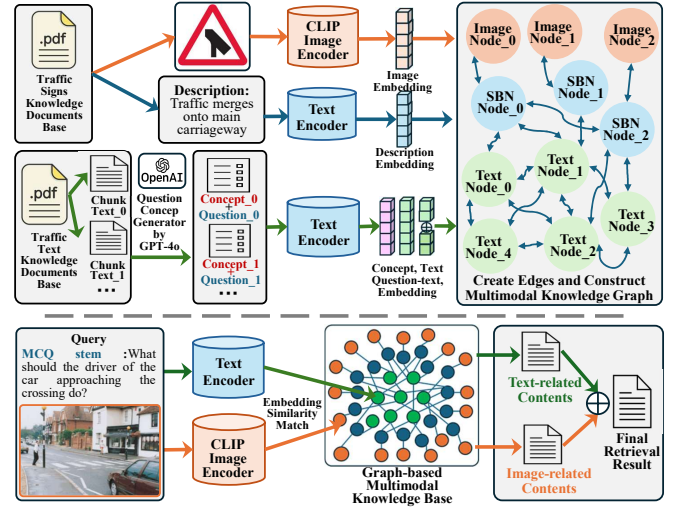


Fig. 2. Overview of the proposed QGM-GRAG approach. The pipeline is divided into two stages: (Top) Knowledge Graph Construction and (Bottom) Retrieval Phase. If a tests does not contain an image, the retrieval stage operates solely on the text input without requiring visual input.

both distributional and lexical relatedness. Image-description edges encode the direct pairing between each traffic sign and its associated description. Description-text edges are weighted by cosine similarity computed in the shared E5 embedding space, enabling cross-modal alignment between visual sign content and regulatory text. This explicit node-edge design ensures structured and interpretable connectivity across modalities within the graph.

This structured representation preserves modality-specific information and supports interpretable, robust cross-modal inference via semantic alignment and multi-hop traversal. Cross-modal alignment is achieved through a two-stage bridging strategy. First, image nodes are matched to description nodes using CLIP-space cosine similarity. Second, description nodes are linked to text nodes via E5-space cosine similarity computed on L2-normalized embeddings. This design avoids projecting image features directly into the text embedding space and instead uses description nodes as semantic bridging anchors, improving stability and interpretability. All text and description embeddings are L2-normalized prior to similarity computation. Edges are added only when similarity exceeds predefined thresholds, and concept-overlap constraints are applied for text-text connections. These filtering strategies reduce spurious connections and mitigate noise propagation during multi-hop retrieval.

B. Multimodal Knowledge Graph Construction

The multimodal knowledge graph is constructed from publicly available UK driving rule documents and official traffic-sign descriptions. Algorithm 1 summarizes the key steps in multimodal knowledge graphs construction. Initially, the source document is segmented into paragraph-level chunks with a maximum length of 1000 characters with 200 overlaps. For each paragraph, GPT-4o extracts 3–5 key concepts and generates 2 to 10 diverse and specific questions that focus

on testing deep understanding of the paragraph's core content. The question generation process is carried out using a structured prompt. The system message instructs the model to act as a question generation assistant for educational reading comprehension, while the user message specifies that the questions must be specific to the key facts, implications, and concepts found in the paragraph.

Both paragraph texts and generated questions are embedded using the multilingual E5 sentence embedding model (intfloat/multilingual-e5-large-instruct, 1024 dimensions). To create a unified representation that captures both the paragraph content and the questions, we use a weighted sum approach to fuse the paragraph embedding, e_t , with the mean question embedding, e_q , as follows: $e_{joint} = \alpha e_t + (1 - \alpha)e_q$, where $\alpha = 0.5$ by default. After fusion, we apply L2 normalization to the joint embedding to ensure consistent similarity computation across different nodes: $e_{joint} = \frac{e_{joint}}{\|e_{joint}\|}$. This joint embedding is then used to represent each text node in the knowledge graph. Edges between text nodes are established based on both semantic similarity between their joint embeddings and lexical overlap of the concepts extracted from each paragraph. Specifically, the weight between two text nodes i and j is computed as

$$w_{ij} = \alpha \cdot \cos(\mathbf{z}_i, \mathbf{z}_j) + (1 - \alpha) \cdot O(C_i, C_j), \quad (1)$$

where $\mathbf{z}_i$ and $\mathbf{z}_j$ denote the joint embeddings of paragraph i and j . In our implementation, the joint embedding is constructed as a weighted sum of the paragraph embedding and the mean question embedding, followed by L2 normalization:

$$\mathbf{z}_i = \frac{\alpha \mathbf{e}_i + (1 - \alpha) \bar{\mathbf{e}}_{Q_i}}{\|\alpha \mathbf{e}_i + (1 - \alpha) \bar{\mathbf{e}}_{Q_i}\|_2}, \quad (2)$$

where $\mathbf{e}_i$ denotes the E5 embedding of paragraph i , and $\bar{\mathbf{e}}_{Q_i}$ is the mean embedding of its generated guiding questions. The lexical overlap term $O(C_i, C_j)$ is defined as a normalized overlap coefficient:

$$O(C_i, C_j) = \frac{|C_i \cap C_j|}{\min(|C_i|, |C_j|)}, \quad (3)$$

where C_i and C_j represent the sets of key concepts extracted from the two paragraphs.

In addition, we integrate visual knowledge by processing traffic-related images. Images and their corresponding textual descriptions are collected from UK official highway code resources. Each image is paired with a brief textual description, which we embed with the same E5 encoder used for paragraphs and store as a description ("desc") node acting as a semantic bridging node (SBN). The image itself is encoded in the CLIP space, and we first retrieve similar image nodes using CLIP-based cosine similarity. We then traverse to their associated desc nodes and finally to text nodes through E5-space edges whose weights are the cosine similarities between the desc embedding and the text joint embeddings. This two-step bridge (image $\rightarrow$ desc in CLIP/E5, then desc(E5) $\rightarrow$ text(E5)) aligns visual content with the textual knowledge graph without projecting image features into the text embedding space. We apply L2 normalization to all embeddings, including those for text, questions, and images, before calculating similarities to

ensure consistent comparisons across the different types of content. In our implementation, each traffic sign image is paired with a curated textual description obtained from the official traffic sign documentation repository. These descriptions are not generated dynamically by LLM but originate from standardized reference materials to ensure semantic consistency and reliability. No additional paraphrasing or rewriting is applied during graph construction.

In our implementation, a text-text edge is created only if w_{ij} exceeds a threshold $\theta_{\text{text}} = 0.6$, which was empirically chosen to balance graph sparsity and semantic coverage. For desc-text edges, a cosine similarity threshold $\theta_{\text{cross}} = 0.5$ is applied. These thresholds were chosen to balance graph sparsity and semantic connectivity.

While the current rule base and dataset are UK-specific, the proposed framework can be adapted to other countries with minimal modifications. To achieve this, the rule corpus would need to be replaced with country-specific driving rules, theory questions, and regulations. The graph construction method itself remains unchanged, but the rule corpus and traffic sign data from the target country would need to be used for embedding and graph creation. Additionally, traffic sign data and descriptions would need to be re-collected from local sources or government databases to ensure they align with the target country's traffic rules and signage system. By adapting these components, the proposed method can be readily extended to other countries or regions, enhancing its generalizability.

C. Question-Guided Multimodal Graph Retrieval

The retrieval procedure, as shown in Fig. 2, supports three entry modes and leverages the graph structure to assemble answer-grounding evidence. Algorithm 2 details the retrieval process. In the retrieval phase, we compute the query embedding, e_q , based on the user input, which can be a text query, question-form query, or image query. For a text query, the query embedding e_q is computed using the same E5 encoder with the prefix "query:". We then compute the cosine similarity between the query embedding e_q and the joint embeddings of the text nodes in the graph, as follows: Cosine Similarity: $\text{sim}(q, t_i) = \cos(e_q, e_{joint}(i))$ where $e_{joint}(i)$ is the joint embedding of text node i . The top-k ($k=3$) nodes are retrieved based on this similarity.

For a question-form query, we compare the query against each node's per-question embeddings. For each text node, a set of per-question embeddings is generated. The query embedding e_q is compared to each of these per-question embeddings, and we retrieve the top-k text nodes based on the maximum per-question similarity, ensuring the most relevant supporting text nodes are retrieved for the given question.

For an image query, the input image is encoded into a CLIP embedding e_x . The cosine similarity between the image query embedding e_x and the image nodes in the graph is computed, and a CLIP threshold ($\theta_{\text{clip}} = 0.9$) is applied to filter out low similarity results. The top-k image nodes are retrieved based on their cosine similarity scores. After selecting the most relevant image nodes, we traverse from the image nodes

Algorithm 1 QGM-GRAG: Multimodal Knowledge Graph Construction

Require: Text corpus T (paragraph chunks with max length 1000 chars and 200 overlap), image collection I , textual descriptions D , thresholds $\theta_{\text{text}}, \theta_{\text{cross}}$, weighting coefficient α

Ensure: Multimodal knowledge graph $G = (V, E)$

- 1: $G \leftarrow \text{INITGRAPH}()$
- 2: *// All text (paragraphs, questions, descriptions) are encoded by E5 with the "query:" prefix and L2-normalization. Image features are from CLIP and L2-normalized.*
- 3: **for each** paragraph chunk $t \in T$ **do**
- 4: $C_t \leftarrow \text{LLM_CONCEPTS}(t)$ *// 3–5 key concepts*
- 5: $Q_t \leftarrow \text{LLM_QUESTIONS}(t)$ *// 2–10 guiding questions*
- 6: $\mathbf{e}_t \leftarrow \text{EMBEDTEXT_E5}(\text{"query: " + } t)$ *// L2-normalized*
- 7: $\mathcal{E}_{Q_t} \leftarrow [\text{EMBEDTEXT_E5}(\text{"query: " + } q)]_{q \in Q_t}$ *// per-question embeddings (L2-norm)*
- 8: $\bar{\mathbf{e}}_{Q_t} \leftarrow \frac{1}{|Q_t|} \sum_{\mathbf{e}_q \in \mathcal{E}_{Q_t}} \mathbf{e}_q$
- 9: $\bar{\mathbf{e}}_{Q_t} \leftarrow \bar{\mathbf{e}}_{Q_t} / \|\bar{\mathbf{e}}_{Q_t}\|_2$
- 10: $\mathbf{z}_t \leftarrow \alpha \mathbf{e}_t + (1 - \alpha) \bar{\mathbf{e}}_{Q_t}$
- 11: $\mathbf{z}_t \leftarrow \mathbf{z}_t / \|\mathbf{z}_t\|_2$
- 12: $v_t \leftarrow \text{ADDTXTNODE}(G, t, \mathbf{z}_t, C_t, \mathcal{E}_{Q_t})$
- 13: **end for**
- 14: **for each** pair of text nodes (v_i, v_j) **do**
- 15: $o_{ij} \leftarrow \frac{|C_{t_i} \cap C_{t_j}|}{\min(|C_{t_i}|, |C_{t_j}|)}$
- 16: $s_{ij}^{\text{cos}} \leftarrow \text{COSINE}(\mathbf{z}_{t_i}, \mathbf{z}_{t_j})$
- 17: $w_{ij} \leftarrow \alpha \cdot s_{ij}^{\text{cos}} + (1 - \alpha) \cdot o_{ij}$ *// $\alpha \in [0, 1]$*
- 18: **if** $w_{ij} > \theta_{\text{text}}$ **then**
- 19: $\text{ADDEDGE}(G, v_i, v_j, \text{type} = \text{"text-text"}, \text{weight} = w_{ij})$
- 20: **end if**
- 21: **end for**
- 22: **for each** image $x \in I$ with paired description $d \in D$ **do**
- 23: $\mathbf{e}_x \leftarrow \text{EMBEDIMAGE_CLIP}(x)$ *// L2-normalized CLIP feature*
- 24: $\mathbf{e}_d \leftarrow \text{EMBEDTEXT_E5}(\text{"query: " + } d)$ *// E5 text space (L2-norm)*
- 25: $v_x \leftarrow \text{ADDIMAGENODE}(G, x, \mathbf{e}_x)$ *// type="image"*
- 26: $v_{\text{desc}} \leftarrow \text{ADDDDESCNODE}(G, d, \mathbf{e}_d)$ *// type="desc" (implements SBN)*
- 27: $\text{ADDEDGE}(G, v_x, v_{\text{desc}}, \text{type} = \text{"image-desc"}, \text{weight} = 1.0)$
- 28: **for each** text node v_t in G **do**
- 29: $s^{\text{cross}} \leftarrow \mathbf{e}_d^\top \mathbf{z}_t$ *// cosine since both are L2-normalized*
- 30: **if** $s^{\text{cross}} > \theta_{\text{cross}}$ **then**
- 31: $\text{ADDEDGE}(G, v_{\text{desc}}, v_t, \text{type} = \text{"desc-text"}, \text{weight} = s^{\text{cross}})$ *// stored for retrieval ranking*
- 32: **end if**
- 33: **end for**
- 34: **end for**
- 35: **return** G

to their associated description (desc) nodes, which serve as semantic bridging nodes (SBN). From the description nodes, we continue to relevant text nodes using edges weighted by the cosine similarity between the description embedding $\mathbf{e}_d$ and the joint embedding of the text nodes.

The retrieval process involves multi-hop expansion over the text, description, and image nodes, where the top-k nodes are ranked by their similarity scores and edge weights. After collecting the relevant nodes, we aggregate the context and use it to generate the final answer. Multi-hop expansion is bounded by a hop budget $H=2$. During expansion, only neighbors whose edge weights exceed their respective thresholds are considered. At each hop, nodes are ranked by edge weight and only the top-k neighbors are retained, preventing excessive propagation

Algorithm 2 QGM-GRAG: Question-Guided Multimodal Retrieval

Require: Graph G , user query q (text/question/image), top- k settings $(k_{\text{text}}, k_q, k_{\text{img}})$, hop budget H , CLIP threshold θ_{clip} (e.g., 0.9)

Ensure: Retrieved context $\mathcal{C}$, final answer a

- 1: $\mathcal{S} \leftarrow \emptyset$ *// entry set*
- 2: **if** $\text{ISTEXTORMCQ}(q)$ **then**
- 3: $\mathbf{e}_q \leftarrow \text{EMBEDTEXT_E5}(\text{"query: " + } q)$ *// L2-normalized*
- 4: $\mathcal{S}_{\text{text}} \leftarrow \text{TOPKTEXTBYJOINTCOSINE}(G, \mathbf{e}_q, k_{\text{text}})$
- 5: $\mathcal{S} \leftarrow \mathcal{S} \cup \mathcal{S}_{\text{text}}$
- 6: **end if**
- 7: **if** $\text{ISQUESTIONFORM}(q)$ **then**
- 8: $\mathbf{e}_q \leftarrow \text{EMBEDTEXT_E5}(\text{"query: " + } q)$
- 9: $\mathcal{S}_q \leftarrow \text{TOPKTEXTBYPERQUESTION}(G, \mathbf{e}_q, k_q)$ *// max over node.question_embeddings*
- 10: $\mathcal{S} \leftarrow \mathcal{S} \cup \mathcal{S}_q$
- 11: **end if**
- 12: **if** $\text{ISIMAGE}(q)$ **then**
- 13: $\mathbf{e}_x \leftarrow \text{EMBEDIMAGE_CLIP}(q)$ *// L2-normalized*
- 14: $\mathcal{I} \leftarrow \text{TOPKIMAGESBYCLIP}(G, \mathbf{e}_x, k_{\text{img}}, \theta_{\text{clip}})$ *// filter by CLIP cosine*
- 15: **for each** image node $v_{\text{img}} \in \mathcal{I}$ **do**
- 16: **for each** neighbor u of v_{img} with type="desc" **do**
- 17: $\mathcal{S} \leftarrow \mathcal{S} \cup \{u\}$ *// desc nodes as bridging entries*
- 18: **end for**
- 19: **end for**
- 20: **end if**
- 21: $\mathcal{N} \leftarrow \text{GRAPHEXPAND}(G, \mathcal{S}, H)$ *// multi-hop over {text, desc, image}; follow edge weights*
- 22: $\mathcal{C} \leftarrow \text{AGGREGATECONTEXT}(\mathcal{N})$ *// rank by entry similarity + edge weights; dedup; length-trim*
- 23: $a \leftarrow \text{LLM_ANSWER}(q, \mathcal{C})$ *// text-only output*
- 24: **return** $\mathcal{C}, a$

of weakly related nodes.

IV. SCENARIO-BASED EVALUATION FRAMEWORK

A. Overview of Proposed Framework

To systematically evaluate the transfer of improved theoretical driving knowledge to real-world decision-making, we introduce a structured scenario-based evaluation framework, as illustrated in Fig. 3. This framework creates realistic driving scenarios derived from official UK driving theory content, enabling principled assessment of LLMs' ability to translate theoretical knowledge into context-sensitive driving decisions.

The framework comprises three main components. (1) Theory-to-scenario transformation: We select MCQs that baseline LLMs fail to answer correctly but answer correctly after enhancement (e.g., RAG or fine-tuning). Each selected question is converted into a detailed textual grounded driving scenario using a GPT-4o-based prompt module, embedding theoretical concepts into realistic and complex traffic contexts. (2) Decision generation: These detailed scenarios textual descriptions, such as road signs or road layouts, are provided to the target LLM. The model must interpret each scenario and generate contextually appropriate driving decisions (e.g., yielding, turning, or proceeding). (3) Automated expert Evaluation: Each decision generated by the LLM is assessed by a GPT-4o-based evaluator trained to emulate expert driving judgment. Decisions are scored on multiple dimensions, such as legal compliance, behavioral safety, risk awareness, and

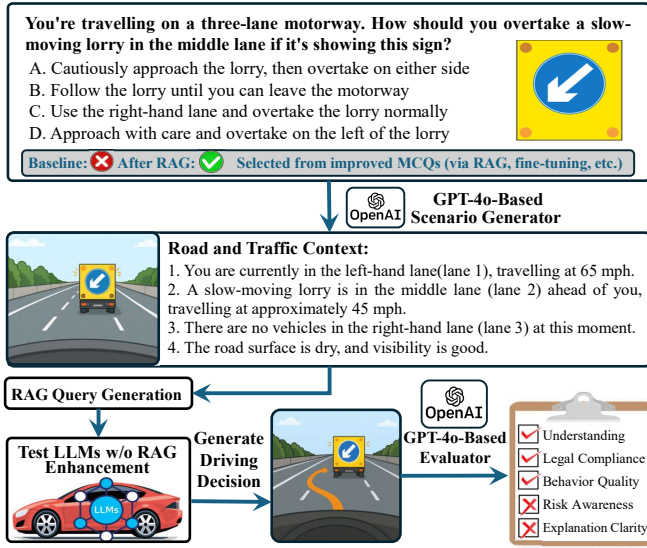


Fig. 3. Overview of proposed scenario-based evaluation framework. Starting from multiple-choice driving theory questions, the framework selects those that baseline LLMs failed to answer but improved after enhancement (e.g., RAG, fine-tuning, Retraining). These are transformed into context-rich scenarios via a GPT-4o-based generator. The enhanced or baseline LLMs are then prompted to generate decisions, which are evaluated by GPT-4o across five dimensions, such as scenario understanding, sign interpretation, legal compliance, driving behavior quality, and risk awareness.

explanation clarity. An example of proposed scenario-based evaluation framework can be found in the Appendix A.

B. Theory-to-Scenario Transformation

The selected MCQs indicate theoretical knowledge gains and form the foundation for scenario generation. We then employ a GPT-4o-based generation module, simulating an expert UK driving instructor familiar with the UK Highway Code and real-world traffic conditions. Each MCQ input comprises four elements: (1) question stem, (2) four multiple-choice options (A–D), (3) the correct answer, and (4) an optional associated image. The prompt guides the model to identify the underlying driving concept (e.g., overtaking procedures, junction priorities, pedestrian awareness) and construct a comprehensive driving scenario around this concept.

The resulting scenarios consistently include five structured components: (1) Road and environment details, which include road type, lane configuration, speed limits, and visibility conditions. (2) Traffic situation, dealing with the positions, behaviors, and movements of surrounding vehicles and road users. (3) Environmental context, dealing with weather conditions, lighting, roadworks, and relevant signage. (4) Driver condition (optional), dealing with internal states such as fatigue or distraction when applicable. (5) High-level driving intention, dealing with clearly stated driver objectives (e.g., “Turn left”, “Overtake another vehicle”) selected from predefined options. To ensure scenario consistency and enable reliable downstream automated evaluation, the GPT-4o module enforces a strictly defined structured output format, thereby eliminating extraneous text. Manual verification on a representative subset of generated scenarios confirmed that GPT-4o consistently

produced valid and contextually coherent transformations in accordance with established traffic regulations. By incorporating GPT-4o as an expert-in-the-loop, the proposed framework enables scalable and high-fidelity conversion of theoretical knowledge points into realistic and evaluable driving contexts. This structured design guarantees that each scenario faithfully encapsulates the original theoretical concept within a practical decision-making environment, thereby facilitating rigorous assessment of theory-to-behavior transfer.

C. Driving Decision Generation

After converting theory-test questions into driving scenarios, we next evaluate whether edge-deployable small-scale LLMs (e.g., Gemma3-4B-it) can effectively interpret and respond to realistic driving situations. To achieve this, we design a structured decision-generation interface that prompts models to provide immediate, safe, and legally compliant driving responses aligned with UK traffic regulations and best practices. The target LLM is explicitly guided to act as a UK-based driving assistant, ensuring strict adherence to the UK Highway Code, compliance with left-hand traffic rules, and prioritization of safety. Our designed prompt enforces a structured, multi-dimensional response comprising five key elements. (1) Situation analysis, which summarizes current road and traffic conditions, clearly identifies hazards, and articulates factors contributing to scenario complexity. (2) Recommended driving actions, which provide clear, sequential driver actions, covering precise maneuvers (lane changes, braking), safety checks (mirrors, signaling), and proper vehicle positioning. (3) Legal and safety justifications, which explicitly refer relevant UK Highway Code rules to justify recommended actions, reinforcing legal accuracy and rule-based decision inference. (4) Risks of incorrect actions, which clearly articulates potential negative consequences of improper or unlawful driving decisions, including collisions, penalties, or risks to vulnerable road users. (5) Enhanced driving advice, which offers proactive driving strategies, such as hazard anticipation and appropriate spacing, demonstrating a comprehensive grasp of driving safety principles. This structured approach accommodates both purely textual LLMs and vision-language models through scenario imagery integration, bridging theoretical knowledge with practical behavioral responses and enabling rigorous downstream evaluation.

D. Automated Expert Evaluation

To assess the quality of generated driving decisions, we implement a structured evaluation pipeline using GPT-4o as an automated expert assessor. The system prompt defines the evaluator’s role as a professional UK driving examiner with expertise in traffic law, highway safety, and behavioral risk assessment. The user prompt enforces a highly structured rubric, guiding the model to assess each response across five distinct dimensions. (1) Scenario understanding, which assesses whether the model correctly interprets the spatial and contextual elements of the scene. (2) Legal compliance, which assesses whether its recommendations adhere to relevant UK Highway Code rules; (3) Driving behavior quality, which

assesses whether the prescribed actions are safe, realistic, and procedurally sound. (4) Risk awareness, which assesses whether potential hazards are identified and mitigated. (5) Explanation clarity, which assesses whether the response is well-structured and accessible for learners.

Each dimension is scored on a 0–2 scale, accompanied by detailed comments. A final overall score on a 0–10 scale and suggested revisions are also generated. Consistency was ensured by using a standardized, rubric-based prompt with fixed evaluation dimensions. To verify robustness, we repeated scoring on a sample subset and observed minimal variance across runs, suggesting reliable reproducibility. This enables scalable, reproducible, and human-like assessment of LLM behavior in realistic driving scenarios. To reduce stylistic bias, the evaluation prompt explicitly instructs the evaluator to focus on decision correctness, safety compliance, and rule adherence rather than answer length, rhetorical structure, or linguistic style. The scoring rubric defines discrete levels based on behavioral appropriateness and legal correctness within the scenario. No stylistic features are considered in scoring. To assess evaluation robustness, we conducted two supplementary checks. First, each response was independently evaluated twice under identical prompt conditions, and we observed minimal score variance (standard deviation < 0.3 on a 5-point scale). Second, we performed a small-scale manual validation on a randomly sampled subset ($n=10$). Human judgments showed an agreement rate of 90% with automated evaluation results, supporting the consistency of the evaluation framework.

In the appendix, we provide a set of examples, including driving decisions generated using Gemma3-4B-it for an example driving scenario, and the evaluation of those decisions across five dimensions using GPT-4o as an evaluation expert. We also conducted multiple rounds of review of the driving decision generation and decision quality assessment process using human driving experts, validating the effectiveness of the evaluation process.

V. EXPERIMENTS AND RESULTS

A. Experiments Settings

To systematically evaluate the impact of retrieval-augmented knowledge on both driving theory test performance and scenario-based driving decision quality, we design a multi-stage experiment pipeline. This section details the datasets, models, and computational environment used throughout our study.

1) *Data Sources*: We construct our evaluation set from a curated collection of publicly available UK driving theory MCQs. These items are sourced from online repositories, learner driver platforms, and sample exams that are likely derived from official DVSA (Driver and Vehicle Standards Agency) materials and commercial driving school databases. Each MCQ consists of a textual driving prompt, four answer options (A–D), and a single correct answer. To support multimodal analysis, the dataset includes both text-only and image-based questions, with the latter presenting photographs or illustrations of road environments, signage, or vehicle positioning.

For a more fine-grained evaluation of RAG performance across different knowledge areas, we manually categorize all questions into seven semantic classes: (1) Driving Behavior & Safe Practice, (2) Road Signs & Markings, (3) Driver Responsibilities & Legal Rules, (4) Hazard Awareness & Emergency Situations, (5) Eco-driving & Vehicle Maintenance, (6) Pedestrians & Other Road Users, and (7) Junctions & Traffic Signals. This classification enables us to assess not only overall model accuracy but also how retrieval augmentation affects different types of driving knowledge and decision-making. Further details on question structure, dataset composition, and classification methodology can be found in our earlier work [8].

2) *Small-scale LLMs*: Our primary focus is on evaluating compact, open-source LLMs suitable for on-vehicle deployment. We consider models in the 1B–10B parameter range, selected for their efficiency and real-time applicability in resource-constrained environments. Table I summarizes the key characteristics of the three compact LLMs used in our study: Gemma3-4B-it [24], Qwen2.5-VL-3B-Instruct, and Qwen2.5-VL-7B-Instruct [25]–[27]. All models are open-source and instruction-tuned, making them suitable for deployment on edge devices or vehicle-mounted platforms. Gemma3-4B-it adopts a decoder-only architecture and focuses on text and image interaction with a significantly extended context length of up to 128k tokens. Both Qwen2.5-VL variants are vision-language transformers capable of processing text and image inputs. These models support long-context understanding (up to 32k tokens), which facilitates the integration of retrieved driving knowledge and enables coherent multi-turn interactions essential for autonomous driving tasks. Throughput values (tokens per second) were measured using FP16 inference on a single NVIDIA A100 80GB GPU under batch size 1 and maximum input length settings, consistent with open-source benchmarks such as those from Hugging Face Open LLM Leaderboard [28]. These metrics reflect realistic performance expectations in latency-sensitive driving applications. We additionally employ GPT-4o in auxiliary roles for chunk indexing, scenario generation, and decision evaluation, as already described in Sections IV-D.

3) *Experimental Environment*: All experiments are conducted in a unified, high-performance computing environment featuring a single NVIDIA H100 PCIe GPU with 40 GB of memory. By standardizing the hardware platform, we eliminate variability due to computational resources and focus purely on model performance and architectural differences. This environment provides sufficient memory and throughput to support both lightweight open-source models and auxiliary use of GPT-4o via API.

4) *Evaluation Metrics*: To comprehensively assess the performance of LLMs under both theory-testing and decision-making conditions, we define two complementary sets of evaluation metrics. For the first stage of evaluation, we measure the accuracy of LLMs on a curated set of UK driving theory MCQs. Accuracy is computed as the proportion of correctly answered questions over the total number of evaluated ques-

TABLE I
OVERVIEW OF LLMs USED IN OUR EXPERIMENTS

Model Name	Parameter Size	Modality	Open Source	Architecture	Max Input Length	Throughput (tokens/s)
Gemma-3-4B-it	4.30B	Text + Image	Yes	Decoder-only	128,000	~950
Qwen2.5-VL-3B-Instruct	3.75B	Text + Image	Yes	Vision-Language Transformer	32,768	~850
Qwen2.5-VL-7B-Instruct	8.29B	Text + Image	Yes	Vision-Language Transformer	32,768	~800

TABLE II
GRAPH AND RETRIEVAL HYPERPARAMETERS

Parameter	Value
Chunk size	1000 chars
Chunk overlap	200 chars
Concepts / paragraph	3–5
Questions / paragraph	2–10
Embedding model	mE5-large (1024-d)
Joint weight α	0.5
Text–text threshold θ_{text}	0.6
Desc–text threshold θ_{cross}	0.5
CLIP threshold θ_{clip}	0.9
Hop limit H	2
Top- k (text)	3
Top- k (question)	5
Top- k (image)	3

TABLE III
COMPARISON OF BASELINE VARIANTS AND THEIR COMPONENTS

Method	Text Node	Question Guidance	Multimodal (SBN)
GRAG	✓	×	×
QG-GRAG	✓	✓	×
QGM-GRAG	✓	✓	✓

and prompting template. These settings guarantee that any performance differences are solely due to the methodologies themselves, rather than variations in the experimental setup. Each LLM’s response was evaluated three times and the mean score was reported to reduce stochastic variance. We also verified that minor paraphrasing of model outputs did not materially affect evaluation scores, indicating that the scoring is content-sensitive rather than phrasing-sensitive.

5) *Adapting to Other Countries*: The proposed method can be adapted to other countries’ traffic rules by replacing the UK-specific rule corpus with data from the target country. For each new region, the traffic sign data and corresponding descriptions would need to be re-collected, either manually or from publicly available databases. We believe this adaptation is straightforward and does not require significant changes to the underlying methodology. Future work will explore the application of this framework to non-UK driving scenarios.

B. Baseline Method

To assess the effectiveness of our proposed QGM-GRAG approach, we compare it against a series of increasingly simplified baselines that systematically remove core components. (1) Baseline 0 (No RAG): The LLMs answer each question using only its pretrained internal knowledge, without any retrieval. This provides a lower bound and highlights the limitations of zero-context inference. (2) Baseline 1 (T-GRAG): A text-only retrieval-augmented setup where each paragraph forms a graph node. Edges are established based on a hybrid score combining embedding similarity and concept overlap. Retrieval is performed by matching the query embedding with chunk embedding. (3) Baseline 2 (QG-GRAG): An extension of GRAG that introduces LLMs-generated questions for each chunk. These questions are embedded and averaged with the original chunk embedding to create a question-text joint representation, enhancing semantic alignment during retrieval. This isolates the effect of question-guided supervision. A concise comparison of T-GRAG, QG-GRAG, and QGM-GRAG is provided in Table III, which clearly illustrates the incremental enhancement logic across the variants. T-GRAG and QG-GRAG already cover the mainstream RAG paradigms

tions:

$$\text{Accuracy} = \frac{\# \text{ Correct Answers}}{\# \text{ Total Questions}} \times 100\% \quad (4)$$

This metric is reported under both baseline and RAG-enhanced settings for each model. Accuracy improvements quantify the degree to which the external knowledge provided by the RAG framework enhances the factual correctness and rule understanding of the target LLMs.

To evaluate the behavioral quality of decisions generated in the second stage (scenario-based driving decision evaluation), we adopt a structured, rubric-based scoring framework. Each generated decision is scored across five dimensions. An overall decision quality score (out of 10) is calculated as the sum of the five sub-scores. Additionally, we compute average scores per dimension across all evaluated scenarios and analyze the distribution of improvements brought by RAG. To quantify the effect of RAG on overall model decision-making performance, we calculate the *relative improvement in average total score* as follows:

$$\text{Improvement (\%)} = \frac{\text{Score}_{\text{RAG}} - \text{Score}_{\text{No RAG}}}{\text{Score}_{\text{No RAG}}} \times 100\% \quad (5)$$

Where, $\text{Score}_{\text{RAG}}$ and $\text{Score}_{\text{No RAG}}$ denote the average total decision scores achieved by each model with and without RAG respectively. This percentage highlights the relative gain in decision-making performance when RAG is applied.

For reproducibility, all graph construction and retrieval hyperparameters are summarized in Table II. In our experiments, we ensured fairness by keeping all experimental settings identical for both the baseline methods and the proposed QGM-GRAG approach, including the retrieval corpus, chunking strategy, embedding model, top- k retrieval size, context length,

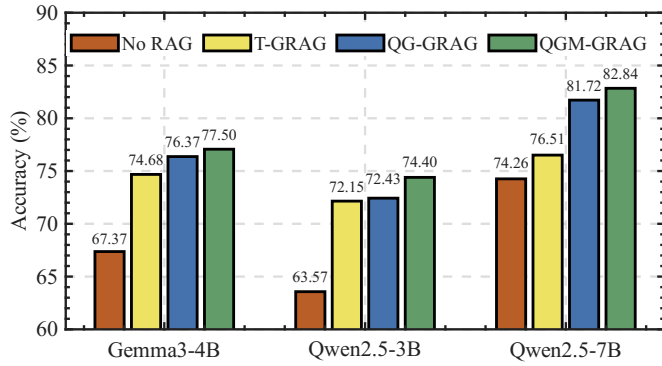


Fig. 4. Overall accuracy comparison across three different LLMs and RAG variants. All models benefit from RAG, with our proposed QGM-GRAG consistently achieving the highest accuracy. The addition of question-guided joint embeddings and SBNs leads to significant performance gains.

widely adopted in recent works. It is important to clarify that the objective of this work is not to replace trajectory-level planning or control-oriented driving systems. Instead, our framework operates at the semantic reasoning layer, aiming to enhance LLM driving-theory competence and quantify its impact on scenario-level decision reasoning. Therefore, comparisons with rule-based or optimization-based control systems are outside the scope of this study, as they operate at different abstraction levels within the driving intelligence stack.

C. Performance on Driving Theory Tests

1) *Overall Accuracy Improvement*: To assess the effectiveness of our proposed QGM-GRAG approach, we evaluate multiple models across four settings: No RAG, T-GRAG, QG-GRAG, and QGM-GRAG. The results, illustrated in Fig. 4, reveal several key observations. First, all tested models achieve significant accuracy improvements when equipped with RAG. Compared to the No RAG baseline, the T-GRAG brings consistent gains across models. For instance, Gemma3-4B-it improves from 67.37% to 74.68%, and Qwen2.5-VL-3B-Instruct increases from 63.57% to 72.43%. This demonstrates that external structured knowledge can substantially enhance a model’s ability to answer driving theory questions. Second, incorporating question-guided representation QG-GRAG further improves retrieval relevance and model performance, particularly in larger models. Qwen2.5-VL-7B-Instruct, for example, rises from 76.51% to 81.72% after introducing question-based joint embeddings. However, for smaller models like Qwen2.5-VL-3B-Instruct, this benefit is less pronounced (e.g., 72.15%→72.43%). This suggests that smaller models may struggle to leverage question-level semantics due to limited capacity.

Finally, our full QGM-GRAG approach consistently achieves the best performance across all models. The integration of multimodal nodes, specifically traffic sign images and their textual descriptions, into the knowledge graph, along with the use of SBNs, results in additional gains beyond question-guided text inference alone. Gemma3-4B-it reaches 77.07%, and Qwen2.5-VL-7B-Instruct achieves 82.84%, outperforming

all baselines. This validates the effectiveness of our design and demonstrates its ability to ground model decisions in both visual and textual knowledge.

In summary, across all categories of driving theory questions, the proposed QGM-GRAG yields higher accuracy than both the no RAG setting and other baselines. These gains are observed uniformly across three compact LLMs, indicating that even relatively small-scale models benefit greatly from the injection of structured external knowledge. The question-guided retrieval strategy clearly enhances alignment between queries and available knowledge, ensuring that the model is provided with precisely the relevant facts or rules needed to answer each question. We also observe that the T-GRAG gain is larger for Qwen2.5-VL-7B than for the 3B model in our setting, likely because the joint question-guided signals align better with the higher-capacity model. While we did not include 14B or 72B models due to resource constraints, we anticipate similar but smaller performance boosts for larger models, with greater relative impact on compact LLMs suitable for edge deployment.

The results from above experiments provide strong evidence for the effectiveness of the QGM-GRAG framework. These experiments validate the impact of our proposed components across different categories of driving theory questions, supporting the main claim of our work: that enhancing driving-theory knowledge improves decision-making. While we acknowledge that additional comparisons with rule-based, learning-based, and optimization-based approaches would provide further context, we argue that the current experiments already demonstrate the core contributions of our framework. Future work could extend this comparison to further validate the general applicability of our approach in the broader context of autonomous driving.

2) *Per-Category Performance Analysis*: To better understand how each retrieval method supports different types of driving knowledge, we present a per-category breakdown of model performance under all four retrieval settings: No RAG, T-GRAG, QG-GRAG, and our full QGM-GRAG method. As shown in Fig. 5, consistent trends emerge across all model scales, reinforcing the robustness of our proposed design.

The most notable improvements are observed in the “Junctions & Traffic Signals” category. This category typically requires multimodal inference over complex visual scenes such as traffic lights, road layouts, and regulatory signs. Our approach introduces SBNs to explicitly link image regions to relevant textual rules, enabling models to accurately retrieve and reason over sign-related knowledge. For example, Qwen2.5-VL-7B-Instruct achieves a substantial improvement from 67.86% (No RAG) to 85.71% with our QGM-GRAG. This highlights the unique advantage of multimodal grounding in scenarios where vision and language need to be jointly interpreted.

Apart from visual-heavy categories, we also observe strong gains in “Hazard Awareness” and “Driver Responsibilities & Legal Rules”. These improvements stem from our question-guided representation mechanism, which aligns user queries with concept-rich content through hybrid embeddings. Even in less visually dependent domains, such as “Eco-driving” and

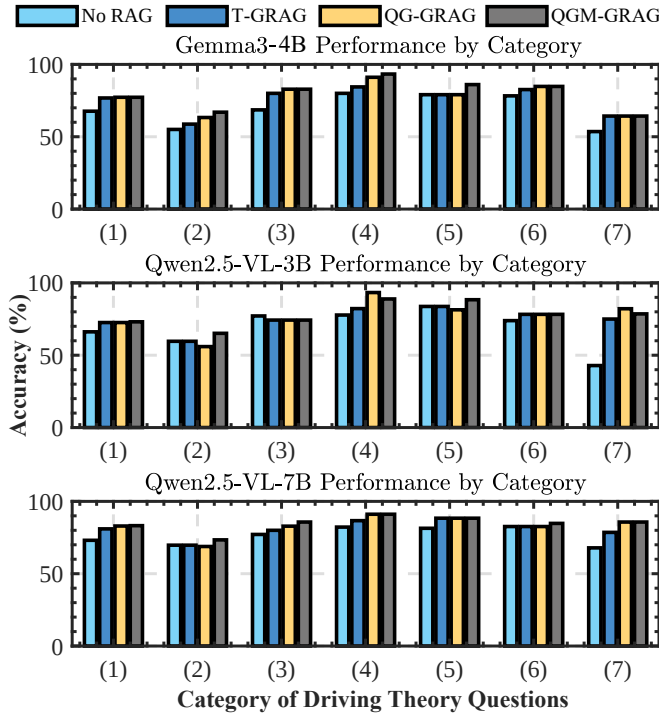


Fig. 5. Per-category accuracy comparison across three LLMs (Gemma3-4B-it, Qwen2.5-VL-3B-Instruct, and Qwen2.5-VL-7B-Instruct) before and after applying our proposed QGM-GRAG. The x-axis represents the 7 categories of driving theory questions: (1) Driving Behavior & Safe Practice, (2) Road Signs & Markings, (3) Driver Responsibilities & Legal Rules, (4) Hazard Awareness & Emergency Situations, (5) Eco-driving & Vehicle Maintenance, (6) Pedestrians & Other Road Users, and (7) Junctions & Traffic Signals. Substantial performance improvements are observed across most categories, highlighting the effectiveness of RAG in supplementing domain-specific driving knowledge.

”Pedestrian Awareness”, the proposed graph-based knowledge retrieval framework helps surface precise and contextually relevant facts, leading to consistent improvements. The benefits of multimodal integration are particularly evident in categories that rely on visual understanding, such as Road Signs & Markings and Junctions & Traffic Signals. In these cases, our approach outperforms all baselines by a larger margin, confirming the value of explicitly linking image features and regulatory text via SBNs. This highlights the necessity of cross-modal knowledge grounding when addressing image-augmented driving questions.

Although QGM-GRAG achieves consistent overall gains, it does not outperform all baselines in every subcategory. In certain visually dominant or straightforward scenarios, the marginal benefit of additional graph retrieval is limited, and excessive retrieval may introduce minor contextual redundancy. In Fig. 5, we observe that the QGM-GRAG approach does not outperform the baseline methods in certain categories when applied to the Qwen2.5-VL-3B model. Specifically, while QGM-GRAG leads to notable improvements in some categories, it shows a slight reduction in accuracy for categories such as Driver Responsibilities & Legal Rules and Junctions & Traffic Signals. This observation may arise from the nature of the Qwen2.5-VL-3B model, which could be less suited for multimodal knowledge integration in specific question

categories. In cases where the model’s baseline performance is already high, the marginal improvement provided by the multimodal knowledge graph may not be as pronounced. Additionally, the potential complexity of integrating multimodal data (text and images) might introduce additional noise or cause overfitting to the specific task in these categories, leading to slightly lower performance. Furthermore, certain categories may require a deeper contextual understanding of real-world scenarios, which the current graph-based retrieval may not fully capture, particularly when compared to simpler textual embeddings used in the baseline models. It is also possible that the threshold settings or weighting factors in the QGM-GRAG model need further tuning to better align with the characteristics of the Qwen2.5-VL-3B model, optimizing performance across all categories.

Overall, these results validate that QGM-GRAG not only excels at multimodal integration but also boosts performance in concept-heavy inference and factual retrieval across diverse domains of driving knowledge.

D. Driving Decision Performance on Generated Scenarios

We filtered for questions that met the following criteria: (1) the baseline model (No RAG) failed to answer them correctly; (2) the same model, when augmented with RAG, produced the correct answer; and (3) the question could be reasonably transformed into a grounded, context-rich driving scenario. Each selected item was then encoded into a detailed driving scenario, incorporating environmental context, road layout, traffic participants, and a clearly defined driving task. For each scenario, we prompted LLMs (with and without our proposed QGM-GRAG pipeline) to generate full-text driving decisions. These outputs were then submitted to GPT-4o, which acted as a consistent evaluator to assess the quality of the responses based on five dimensions. This scenario-to-decision pipeline allowed us to systematically test whether improvements observed at the theory-test level meaningfully translated into more accurate, safer, and law-abiding driving decisions under realistic conditions. The evaluation scenarios cover diverse driving categories, including regulatory reasoning, junction behavior, safety-critical decisions, and multimodal interpretation, mitigating severe intra-domain distributional bias within the studied setting.

Table IV summarizes results from three compact models, reporting average decision scores and fine-grained sub-aspects with and without the support of our QGM-GRAG knowledge retrieval. We observe that all three LLMs experience consistent performance gains across the board when equipped with our RAG method. For instance, the average decision score for Qwen2.5-VL-3B-Instruct improves significantly from 7.29 to 9.00, showing the strongest uplift (+1.71), followed by Qwen2.5-VL-7B-Instruct from 8.15 to 9.05, and Gemma3-4B-it from 8.26 to 8.79. This confirms the utility of our retrieval strategy in supporting not only factual accuracy but also complex inference and judgment in simulated driving scenarios. In terms of sample-wise change statistics, the majority of driving decisions were improved after QGM-GRAG enhancement: 12 out of 17 cases for Qwen2.5-VL-3B-Instruct,

TABLE IV
SUMMARY OF DRIVING DECISION EVALUATION RESULTS FOR THREE COMPACT LLMs.

Model	RAG	#Samples	Improved	Degraded	Unchanged	Avg Score	Scenario	Legal	Behavior	Risk	Explanation
Gemma3-4B	No	19	–	–	–	8.26	1.74	1.16	1.58	1.89	1.89
	Yes	19	9	4	6	8.79	2.00	1.32	1.47	2.00	2.00
Qwen2.5-VL-3B	No	17	–	–	–	7.29	1.76	0.76	1.24	1.76	1.76
	Yes	17	12	4	1	9.00	2.00	1.47	1.71	1.94	1.88
Qwen2.5-VL-7B	No	20	–	–	–	8.15	1.85	1.15	1.30	1.90	1.95
	Yes	20	11	2	7	9.05	1.95	1.55	1.70	1.95	1.95

and 11 out of 20 for Qwen2.5-VL-7B-Instruct. Importantly, the number of degraded answers remains minimal, i.e., fewer than 4, demonstrating the robustness and non-intrusiveness of our framework. Our qualitative analysis shows that, even when the retrieved knowledge is only partially relevant, the models remain robust and their decision quality does not noticeably degrade. The number of evaluated scenarios (17–20 per model) corresponds to the subset of MCQs where (1) baseline models failed but improved with RAG, and (2) the questions could be transformed into realistic scenarios. While this leads to a relatively small sample size, it directly targets the theory-to-scenario transfer effect. Expanding the scenario pool with a larger and more diverse set of MCQs will be an important direction for future work.

Looking into the sub-aspect scores, we note especially strong improvements in the following aspects. (1) Legal reasoning: all models demonstrate a clear jump in their ability to identify the legality of a driving action. For Qwen2.5-VL-3B-Instruct, the legal reasoning score nearly doubles (from 0.76 to 1.47) and rises from 1.15 to 1.55 for Qwen2.5-VL-7B-Instruct. (2) Scenario understanding: QGM-GRAG enables the models to better ground their predictions in the situational context, with Gemma3-4B-it improving from 1.74 to 2.00 and Qwen2.5-VL-3B-Instruct from 1.76 to 2.00. (3) Explanation quality: the explanation sub-score reaches 2.00 for Gemma and 1.95–2.00 for Qwen2.5-VL-7B-Instruct, indicating that the retrieved context supports not only correct decisions but also logically structured and interpretable justifications.

Fig. 6 reports the scenario-level decision performance under No RAG and RAG configurations across four LLM backbones. For each model, the average total score is computed over paired driving scenarios, and 95% confidence intervals are estimated via paired bootstrap resampling at the question level. In addition, the relative improvement ($\Delta\%$) is derived from paired bootstrap samples to ensure statistical consistency between mean estimation and improvement inference.

Across all models, integrating QGM-GRAG consistently yields positive performance gains. Notably, the magnitude of relative improvement exhibits a clear capacity-dependent trend. The smaller-capacity Gemma3-1B-it model demonstrates the largest relative gain, indicating that structured knowledge retrieval substantially compensates for limited internal parameterized knowledge. The mid-scale Qwen2.5-VL-3B-Instruct model also shows statistically significant improvement, suggesting that retrieval augmentation enhances

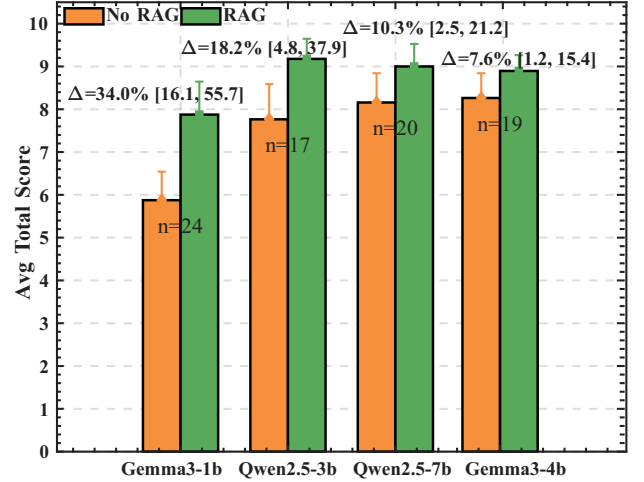


Fig. 6. Comparison of average total decision scores between No RAG and RAG configurations across four LLM backbones. Bars represent mean performance over paired driving scenarios, with error bars indicating 95% confidence intervals estimated via paired bootstrap resampling. The annotated Δ values denote the relative improvement of RAG over No RAG with corresponding 95% confidence intervals. Across all models, integrating QGM-GRAG consistently improves scenario-level decision performance, with the strongest relative gain observed in smaller-capacity models and stable positive gains maintained in larger models.

reasoning alignment even when base capability is moderate. For larger-capacity models such as Qwen2.5-VL-7B-Instruct and Gemma3-4B-it, improvements remain positive but comparatively smaller, implying that while these models already encode substantial driving knowledge, external structured retrieval still refines decision consistency and reduces residual reasoning errors.

Importantly, the paired bootstrap confidence intervals for $\Delta\%$ remain strictly above zero across all backbones, providing statistical evidence that the observed gains are not attributable to sampling variance. This confirms that theory-level knowledge enhancement via QGM-GRAG translates into measurable and statistically reliable improvements in scenario-level decision behavior. Together, these results substantiate the existence of a stable theory-to-decision transfer effect across different model scales.

Fig. 7 presents a radar chart comparing the average performance of the three LLMs across all five evaluation dimensions, both with and without our proposed QGM-GRAG retrieval enhancement pipeline. The chart clearly illustrates the effectiveness of RAG integration across models, with visible

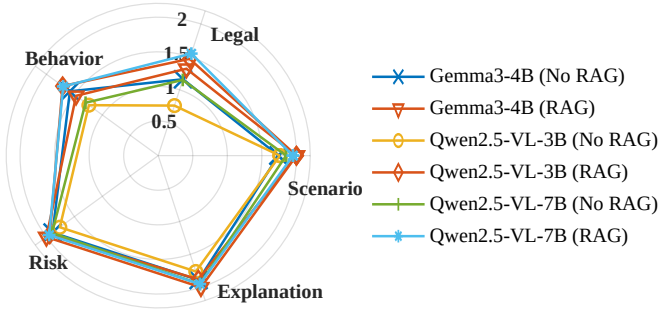


Fig. 7. Radar chart comparison of driving decision quality across five evaluation dimensions for three LLMs, under conditions with and without our proposed QGM-GRAG retrieval augmentation. Each axis represents the average score (0–2 scale) across 20 generated scenarios. The chart shows that all models benefit from RAG, with the most consistent improvements observed in Legal Compliance and Driving Behavior. The larger Gemma3-4B-it model achieves near-maximum performance across all dimensions with RAG support.

radial expansion across all axes. The most pronounced gains are observed in Legal Compliance and Driving Behavior, indicating that retrieval-based access to structured knowledge significantly enhances models’ alignment with statutory rules and appropriate action planning. For instance, the Gemma3-4B-it model exhibits near-saturation performance with RAG support, approaching the upper bound of the 0–2 evaluation scale in nearly all categories. Among the evaluated models, Qwen2.5-VL-3B-Instruct displays the largest performance delta pre- and post-RAG, suggesting that smaller models with strong vision-language capabilities particularly benefit from external knowledge injection. Qwen2.5-VL-7B-Instruct, although already stronger in baseline form, also achieves clear improvements across Legal and Risk dimensions. These trends validate the generality of RAG enhancement and highlight that models with different architectures and parameter sizes can all benefit from retrieval-based decision generation.

The above experimental results underscore the effectiveness and practical utility of our proposed scenario-based evaluation framework in quantifying theory-to-decision transfer within LLMs. The broad yet nuanced improvements observed across multiple evaluation dimensions highlight the sensitivity and precision of the framework, enabling it to capture both overarching performance patterns and subtle variations specific to different models. Consequently, this validates its versatility and applicability across diverse model sizes and theoretical enhancement approaches. By situating evaluations within realistic driving scenarios, our framework successfully elucidates meaningful improvements in practical decision-making capabilities. Although our study specifically demonstrates retrieval-augmented generation as an effective knowledge-enhancement strategy, the general design of this evaluation methodology ensures seamless applicability to other theoretical enhancement methods, such as fine-tuning or knowledge graph integration. As a robust and scalable evaluation paradigm, this framework can readily extend to various real-world decision-making tasks beyond autonomous driving, making it an invaluable asset for advancing future research in embodied and action-oriented AI systems.

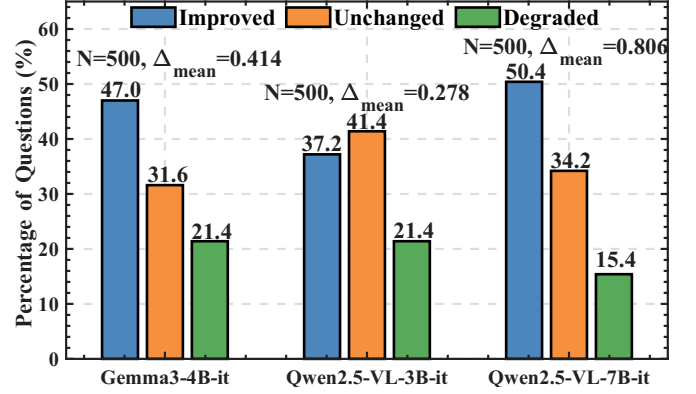


Fig. 8. Outcome distribution of answer changes on the LingoQA dataset after applying the QGM-GRAG pipeline. For each model ($N=500$), bars represent the percentage of questions whose scores are improved, unchanged, or degraded relative to the baseline. The mean score difference (Δ_{mean}) is annotated above each model group to indicate the overall magnitude of improvement.

Despite the general trend, we also observe several counterexamples where models correctly answer theory questions but fail to generate high-quality decisions. Such cases indicate that knowing the rule is not equivalent to applying the rule, underscoring the nonlinear and partial transferability between declarative knowledge and procedural decision inference. For example, while scenario understanding and explanation clarity remain relatively high across models, legal compliance and risk awareness are more dependent on knowledge support, justifying the need for domain-grounded augmentation like RAG. In summary, our proposed method effectively establishes a bridge between static knowledge (theory tests), dynamic inference (scenario interpretation), and practical action (driving behavior generation). RAG not only reinforces memory recall but also facilitates more context-aware and rule-compliant behavior synthesis.

E. Evaluation on LingoQA Dataset

While previous evaluations focused on theory-derived scenarios, we further verify whether the proposed RAG pipeline remains effective in realistic driving contexts. To this end, we conduct experiments on the LingoQA dataset [29], which contains naturalistic driving-related QA pairs. We evaluate three compact LLMs on a subset of LingoQA queries. Our input consists of five consecutive frames of real driving scenes provided by the LingoQA dataset, together with the associated questions. The output is the answer generated by the LLM. We compare the baseline setting with the RAG-enhanced setting, where the retrieved contextual knowledge is used as additional input to support the model’s response, to examine whether retrieval improves answer accuracy. The query used for RAG retrieval is derived from descriptions of the real driving scenes, following the same strategy used for the experiments with the Driving Decision Performance on Generated Scenarios. To account for semantic differences in the responses, we employ GPT-4o as an evaluator, which scores each answer along the dimensions of accuracy and completeness. To ensure consistency and robustness, we repeat the evaluation five

times and report the averaged results. The experimental results are illustrated in Fig. 8. Instead of reporting only overall improvement ratios, we analyze the full distribution of score changes, including improved, unchanged, and degraded cases across 500 questions for each model. All three models exhibit a substantial proportion of improved answers after applying the QGM-GRAG pipeline. Gemma3-4B-it improves on 47.0% of the questions, Qwen2.5-VL-3B-Instruct improves on 37.2%, and Qwen2.5-VL-7B-Instruct achieves the highest improvement ratio at 50.4%. Importantly, the proportion of degraded cases remains comparatively limited across all models, particularly for Qwen2.5-VL-7B-Instruct, indicating that retrieval augmentation does not introduce significant instability.

Beyond improvement frequency, the magnitude of gains further differentiates model behaviors. The larger Qwen2.5-VL-7B-Instruct exhibits the highest mean score increase ($\Delta_{\text{mean}} = 0.806$), substantially exceeding that of the 3B variant ($\Delta_{\text{mean}} = 0.278$). This suggests that stronger backbone models are better able to leverage structured retrieval signals to amplify reasoning quality rather than merely correcting isolated errors. Overall, the distribution-level analysis confirms that the proposed QGM-GRAG pipeline consistently enhances answer quality on realistic driving QA tasks, while maintaining controlled degradation rates, thereby demonstrating stable generalization beyond theory-based evaluation.

VI. CONCLUSION

In this paper, we proposed QGM-GRAG, a question-guided multimodal retrieval-augmented generation framework designed to enhance large language models' driving-theory competence and to evaluate its downstream impact on decision-making. By integrating structured regulatory text and traffic-sign semantics into a unified knowledge graph, the framework improves alignment between natural-language queries and grounded driving knowledge. We further introduced a scenario-based evaluation protocol that enables systematic measurement of theory-to-decision transfer in LLM-based driving systems.

Experimental results demonstrate that QGM-GRAG improves driving-theory test accuracy by over 10% and enhances scenario-level decision performance by more than 13%. Evaluation on the LingoQA dataset further confirms the robustness of the framework in realistic task settings. These findings provide empirical evidence that strengthening factual and procedural driving knowledge can lead to measurable improvements in decision inference quality.

Despite the demonstrated gains, limitations remain. The current knowledge graph is constructed from UK-specific regulatory materials, and the graph structure is static rather than dynamically adaptive. Moreover, evaluation relies on LLM-based automated assessment rather than large-scale human validation. Addressing these limitations will be critical for improving cross-regional transferability, dynamic environmental integration, and evaluation reliability in future work.

VII. FUTURE WORKS

To address the limitations identified above and extend the applicability of the proposed framework, several research

directions can be pursued. First, the framework can be adapted to multi-regional regulatory systems by reconstructing the rule corpus and traffic-sign semantics for different countries. This would allow systematic investigation of cross-domain transfer and regional adaptability. Second, the multimodal knowledge graph can be extended to incorporate dynamic sensory information, such as real-time traffic conditions, weather factors, or infrastructure signals. Integrating such contextual signals would enable more realistic and environment aware decision modeling. Third, future work may explore mechanisms for incremental updating and continual learning, allowing the knowledge graph to evolve based on newly observed driving data. Coupling this with hybrid human-automatic evaluation protocols could further improve evaluation reliability and reduce potential bias. These directions aim to improve scalability, adaptability, and methodological rigor, ultimately advancing the integration of LLM-based reasoning into practical driving intelligence systems.

ACKNOWLEDGMENT

This work was supported in part by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie Actions (MSCA) [Grant Nos. 824019 and 101022280]; in part by the Horizon Europe MSCA Programme [Grant No. 101086228]; in part by the UK Research and Innovation (UKRI) through the Engineering and Physical Sciences Research Council (EPSRC) [Grant No. EP/Y027787/1]; in part by the UKRI EPSRC Researcher Co-Investigator Programme (RCP) [Grant No. 15831/DCM4480]; in part by the National Natural Science Foundation of China with Grants 52571413, 62402353.

REFERENCES

- [1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, and I. Akkaya, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
- [2] C. Cui, Y. Ma, X. Cao, W. Ye, and Y. Zhou, "A survey on multimodal large language models for autonomous driving," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 958–979.
- [3] L. Gan, W. Chu, G. Li, X. Tang, and K. Li, "Large models for intelligent transportation systems and autonomous vehicles: A survey," *Advanced Engineering Informatics*, vol. 62, p. 102786, 2024.
- [4] Y. Li, K. Katsumata, E. Javanmardi, and M. Tsukada, "Large language models for human-like autonomous driving: A survey," in *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2024, pp. 439–446.
- [5] Y. Zheng, Y. Chen, B. Qian, and X. Shi, "A review on edge large language models: Design, execution, and applications," *ACM Computing Surveys*, vol. 57, no. 8, pp. 1–35, 2025.
- [6] R. Qin, D. Liu, C. Xu, Z. Yan, and Z. Tan, "Empirical guidelines for deploying llms onto resource-constrained edge devices," *ACM Transactions on Design Automation of Electronic Systems*, 2024.
- [7] D. Pei, Y. Wu, J. He, and K. Liu, "Methodology and benchmark for automated driving theory test of large language models," *IEEE Transactions on Intelligent Transportation Systems (early access)*, 2025.
- [8] Z. Tang, J. He, D. Pei, and L. Kezhong, "Test large language models on driving theory knowledge and skills for connected autonomous vehicles," in *Proceedings of the 19th Workshop on Mobility in the Evolving Internet Architecture*, 2024, pp. 1–6.
- [9] D. Pei, J. He, K. Liu, and M. Chen, "Application of large language models and assessment of their ship-handling theory knowledge and skills for connected maritime autonomous surface ships," *Mathematics*, vol. 12, no. 15, p. 2381, 2024.
- [10] B. McKinzie, Z. Gan, J.-P. Fauconnier, and S. Dodge, "Mm1: methods, analysis and insights from multimodal llm pre-training," in *European Conference on Computer Vision*. Springer, 2024, pp. 304–323.

- [11] X.-K. Wu, M. Chen, W. Li, and R. Wang, "Llm fine-tuning: Concepts, opportunities, and challenges," *Big Data and Cognitive Computing*, vol. 9, no. 4, p. 87, 2025.
- [12] T. Cai, Y. Liu, Z. Zhou, and H. Ma, "Driving with regulation: Interpretable decision-making for autonomous vehicles with retrieval-augmented reasoning via llm," *arXiv preprint arXiv:2410.04759*, 2024.
- [13] X. Huang, E. M. Wolff, P. Vernaza, and T. Phan-Minh, "Drivegpt: Scaling autoregressive behavior models for driving," *arXiv preprint arXiv:2412.14415*, 2024.
- [14] X. Zhou, X. Han, F. Yang, Y. Ma, and A. C. Knoll, "Opendrivevla: Towards end-to-end autonomous driving with large vision language action model," *arXiv preprint arXiv:2503.23463*, 2025.
- [15] H. Shao, Y. Hu, L. Wang, and G. Song, "Lmdrive: Closed-loop end-to-end driving with large language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 120–15 130.
- [16] L. Halilaj, I. Dindorkar, J. Lüttin, and S. Rothermel, "A knowledge graph-based approach for situation comprehension in driving scenarios," in *European Semantic Web Conference*. Springer, 2021, pp. 699–716.
- [17] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, J. Beißwenger, P. Luo, A. Geiger, and H. Li, "Drivelm: Driving with graph visual question answering," in *European conference on computer vision*. Springer, 2024, pp. 256–274.
- [18] H. Ye, M. Qi, Z. Liu, L. Liu, and H. Ma, "Safedriverag: Towards safe autonomous driving with knowledge graph-based retrieval-augmented generation," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 11 170–11 178.
- [19] M. Arslan, H. Ghanem, S. Munawar, and C. Cruz, "A survey on rag with llms," *Procedia Computer Science*, vol. 246, pp. 3781–3790, 2024.
- [20] S. Kim, "Medbiolm: Optimizing medical and biological qa with fine-tuned large language models and retrieval-augmented generation," *arXiv preprint arXiv:2502.03004*, 2025.
- [21] N. Wiratunga, R. Abeyaratne, L. Jayawardena, K. Martin, and S. Massie, "Cbr-rag: case-based reasoning for retrieval augmented generation in llms for legal question answering," in *International Conference on Case-Based Reasoning*. Springer, 2024, pp. 445–460.
- [22] R. Vasaniya, M. Visodiya, and A. K. Patel, "Techassist: A rag-based chatbot for accessing technical information from stackoverflow," in *2025 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECs)*. IEEE, 2025, pp. 1–6.
- [23] X. Luo, C. Liu, F. Ding, and F. Yang, "Senserag: Constructing environmental knowledge bases with proactive querying for llm-based autonomous driving," in *Proceedings of the Winter Conference on Applications of Computer Vision*, 2025, pp. 989–996.
- [24] A. Kamath, J. Ferret, S. Pathak, N. Vieillard, and R. Merhej, "Gemma 3 technical report," *CoRR*, 2025.
- [25] Q. Team, "Qwen2.5-vl," January 2025. [Online]. Available: <https://qwenlm.github.io/blog/qwen2.5-vl/>
- [26] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, and J. Bai, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.
- [27] J. Bai, S. Bai, S. Yang, S. Wang, and S. Tan, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," *arXiv preprint arXiv:2308.12966*, 2023.
- [28] H. Face, "Open llm leaderboard," https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard, 2024, accessed: May 2025.
- [29] A.-M. Marcu, L. Chen, J. Hünemann, A. Karnsund, B. Hanotte, P. Chidananda, S. Nair, V. Badrinarayanan, A. Kendall, J. Shotton *et al.*, "Lingoqa: Visual question answering for autonomous driving," in *European Conference on Computer Vision*. Springer, 2024, pp. 252–269.

APPENDIX

A. An example of decision-making and scenario-based evaluation framework

We selected a MCQ that models originally answered incorrectly but answered correctly after RAG augmentation. Each such question was then transformed into a detailed driving scenario, simulating realistic road environments, traffic conditions, and decision-making tasks. As illustrated in Fig. 9, we extracted all relevant information, such as road layout, weather, visibility, positions and speeds of vehicles, and required driver actions, and reformulated it into a structured prompt following UK driving norms. This structured prompt was then passed to the model, which was asked to generate a full decision plan for how to handle the situation safely and legally. These scenarios aim to closely mirror the types of complex, multi-variable decisions required in real-world driving contexts.

The models were tasked with analyzing each scenario and generating a safe and legal driving decision plan following UK Highway Code standards. Fig. 10 presents the response generated by Gemma3-4B-it for the scenario described in Fig. 9. The model receives a detailed scene involving a right-turn attempt from a minor road onto a dual carriageway, with clear visibility, known vehicle positions, and relevant signage. In its response, the model interprets the situation, references traffic rules, proposes a step-by-step overtaking plan, and outlines potential risks. While the structure and language are clear and professional, the decision plan includes critical flaws, most notably, the incorrect interpretation of the mandatory left-deviation sign and the illegal suggestion to overtake from the left. These issues highlight the importance of rigorous legal grounding in scenario-based decision-making.

To objectively assess the generated decisions, we employed a standardized evaluation rubric covering five critical dimensions: (1) Scenario Understanding, (2) Legal Compliance, (3) Driving Behavior Quality, (4) Risk Awareness, and (5) Explanation Clarity. Each dimension was scored on a 0–2 scale by an independent evaluation model (GPT-4o), which followed a standardized rubric designed to reflect UK Highway Code compliance and safe driving behavior. Table V presents the evaluation result for the decision generated by Gemma3-4B-it in response to the scenario in Fig. 10. The model earned full marks in scenario understanding and risk awareness, indicating its ability to recognize road structure and hazard presence. However, it received partial scores in three critical dimensions: it misclassified the mandatory directional sign as a general warning triangle, incorrectly suggested overtaking on the left side, and cited an unrelated Highway Code rule.

To evaluate whether improvements in theory-level understanding translate into enhanced practical decision-making capabilities, we conducted a targeted scenario-based assessment. Since not all multiple-choice driving theory questions are suitable for conversion into realistic traffic scenarios, we applied a selection process to identify a subset of questions with meaningful situational structure.

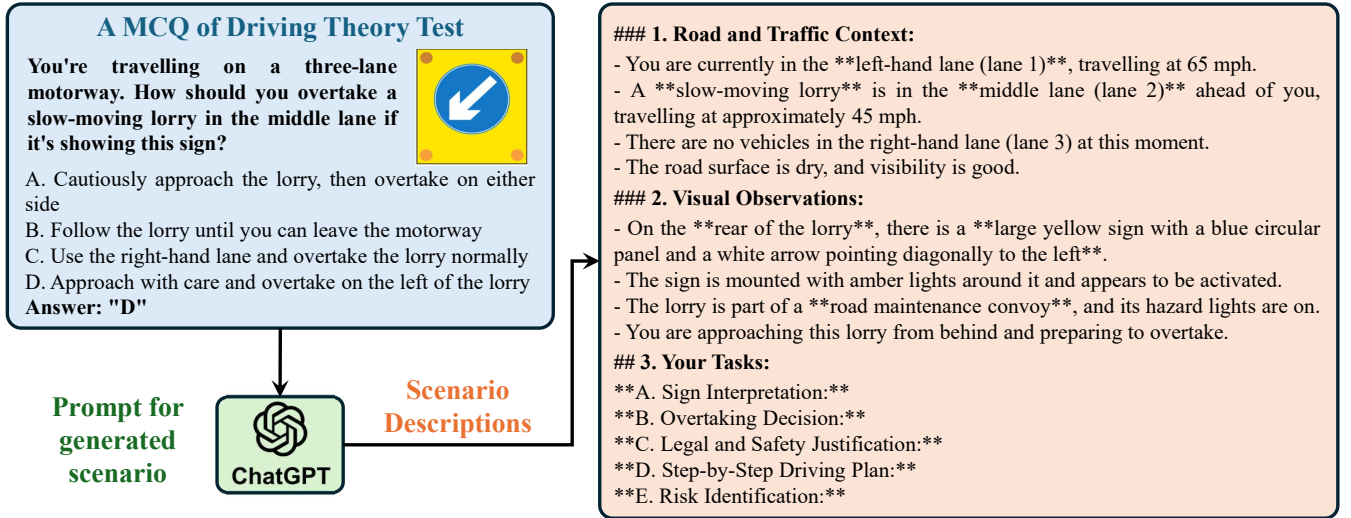


Fig. 9. An example of our MCQ-to-scenario transformation process.

TABLE V
EVALUATION OF THE DRIVING DECISION GENERATED BY GEMMA-3-4B-IT, USING GPT-4O AS THE EVALUATION MODEL.

Evaluation Category	Score (0–2)	Comments
Scenario Understanding	1.5	The model understands the scenario but incorrectly interprets the lorry's sign. In the UK, overtaking on the left is generally not allowed unless in specific circumstances, such as when traffic is moving in queues.
Legal Compliance	0.5	The model misquotes the Highway Code. Rule 18 and Rule 20 do not exist as described. The advice to overtake on the left is incorrect under normal conditions.
Driving Behaviour Quality	1	The sequence of actions is mostly correct, but the recommendation to overtake on the left is unsafe and not compliant with UK driving laws.
Risk Awareness	1	The model identifies some risks but fails to recognize the legal risk of overtaking on the left.
Explanation Clarity	1	The explanation is generally clear but contains incorrect legal advice, which could mislead a learner driver.
Overall Verdict	5.5	The response demonstrates a moderate understanding of the scenario but contains significant legal inaccuracies, particularly regarding overtaking on the left. The advice could lead to unsafe driving practices.

1. Sign Interpretation:

- (1) The sign on the rear of the lorry is a **Warning Triangle**. The yellow background with the blue circular panel and white arrow pointing diagonally to the left indicates a **deviation to the left ahead**. Specifically, it warns of a lane closure or a change in road layout – likely due to road maintenance work. The amber lights and activated hazard lights confirm this is a genuine warning.
- (2) This sign requires following drivers to **reduce speed and be prepared to change lanes to the left** as instructed. It's crucial to anticipate a change in the road ahead.

2. Overtaking Decision:

- (1) Given the three-lane motorway, and the lorry's position in the middle lane, **you should overtake from the right-hand lane (lane 3)**.
- (2) It is **never acceptable** to overtake a slow-moving vehicle on the left in the UK. This is a fundamental rule of motorway driving and a significant safety hazard. Attempting to overtake on the left would be a serious breach of the Highway Code and could lead to a collision.

3. Legal and Safety Justification:

- (1) The primary UK Highway Code rule applicable here is **Rule 118 – Overtaking**. Specifically, Rule 118.1 states: "You must not overtake on the left." Furthermore, Rule 118.2 states: "You must overtake on the right, unless you are overtaking a vehicle going into a rounded on-slip road or a lane."
- (2) It's especially important to follow the sign's instruction **before** attempting any overtaking maneuver. Ignoring the warning sign and overtaking on the left would be a direct contravention of the Highway Code and could lead to a serious accident. The lorry is signaling a potential hazard – you need to react appropriately.

4. Step-by-Step Driving Plan:

- (1) **Observation & Initial Slowdown (Immediate):** Maintain your current speed (65 mph) but begin a gentle, controlled deceleration. Observe the lorry and the road ahead carefully.
- (2) **Signal (Within 200m):** Activate your right-hand turn signal well in advance (at least 200 meters) to indicate your intention to overtake.
- (3) **Lane Change (100m):** Move smoothly and deliberately into the right-hand lane (lane 3). Ensure you have sufficient space and visibility before completing the lane change.
- (4) **Speed Control (100m - Overtake):** Gradually reduce your speed to a point where you can comfortably pass the lorry while maintaining a safe distance. Do not accelerate aggressively. A speed of around 50-55 mph might be appropriate, depending on the road conditions and the lorry's speed.
- (5) **Overtake (100m - 200m):** Pass the lorry with a clear margin of safety. Maintain a consistent speed and position within the right-hand lane.
- (6) **Return to Lane 1 (Immediately After Overtake):** As soon as you are safely past the lorry, signal left and move smoothly back into the left-hand lane (lane 1).
- (7) **Resume Speed (Once in Lane 1):** Gradually accelerate back to your previous speed of 65 mph, maintaining a safe distance from the lorry and any following vehicles.

5. Risk Identification:

- (1) **Overtaking on the wrong side (left):** This is the most significant risk. It would create a direct collision hazard with the lorry and any vehicles behind it.
- (2) **Debris:** Road maintenance work often involves loose materials (gravel, tarmac) which could be dislodged by the road crew and pose a hazard to passing vehicles.
- (3) **Following Maintenance Vehicles:** Other vehicles in the road maintenance convoy could be travelling at slower speeds or making unexpected movements.
- (4) **Reduced Visibility:** Although visibility is currently good, it could deteriorate quickly due to the road work.
- (5) **Sudden Lane Changes:** The lorry driver may suddenly change lanes, requiring you to react quickly.

Fig. 10. Driving decision response generated by Gemma3-4B-it for the scenario shown in Fig. 9.