CarDrivingTutor - A Multi-Modal Framework for Answering Car Driving Exam Questions

Stefan-Paul Buracu
National University of Science and Technology POLITEHNICA Bucharest
313 Splaiul Independetei,
Bucharest, Romania
buraku.stefan@gmail.com

Laurentiu-Marian Neagu
National University of Science and Technology POLITEHNICA Bucharest
313 Splaiul Independetei,
Bucharest, Romania
laurentiu.neagu@upb.ro

Mihai Dascalu
National University of Science and Technology POLITEHNICA Bucharest
313 Splaiul Independetei,
Bucharest, Romania
mihai.dascalu@upb.ro

ABSTRACT

Successful preparation for driving license examinations requires accurately interpreting both textual information and visual cues to represent realistic driving situations. Current systems for answering driving license exam questions in Romania predominantly rely on textual understanding. This paper introduces CarDrivingTutor - a multimodal framework for answering multiple-choice driving license questions in Romania, integrating a Retrieval-Augmented Generation (RAG) strategy using Chroma DB with semantic search, and YOLOv8 for road sign detection. To study the knowledge of LLMs in this area, two models, Gemma3:27B and LLaVA:34B, are analyzed under 3 scenarios ranging from a simple version where the model receives only the question and the answer options, a version in which each question is accompanied by relevant legal context retrieved from the Road Safety Legislation through semantic search, and a version where the textual context from the previous point is also accompanied by official definitions extracted by automatic detection of traffic signs from images. Results show that added textual context significantly improves accuracy, especially for Gemma3. Its accuracy increased from 52% to 64% for questions without images. In contrast, LLaVA showed smaller gains, from 46% to 51%. The use of visual definitions improved Gemma3's accuracy slightly (from 46% to 51%), while LLaVA's performance remained unchanged.

Author Keywords
Multi-modal Framework; Driving License Exam; Large Language Models; Retrieval-Augmented Generation; Semantic Search; Traffic Sign Recognition; YOLO

ACM Classification Keywords
I.2.7 Natural Language Processing; I.4.8 Scene Analysis; K.3.1 Computer Uses in Education.

General Terms
Design; Measurement; Experimentation.

DOI: 10.37789/icusi.2025.13

INTRODUCTION

Large Language Models (LLMs) have become increasingly relevant in educational applications, enabling personalized learning, automated assessments, and enhanced access to domain-specific knowledge [13]. LLMs such as ChatGPT, Gemma, or DeepSeek have advanced capabilities in understanding and interpreting natural language, making them suitable for tutoring, question answering, and even automated grading. Combined with multimodal frameworks, these models can process not only textual data but also images or diagrams, enabling more comprehensive responses in educational settings.

Despite this potential, many educational domains still rely on static or non-interactive resources. In Romania, for example, the theoretical driving license exam includes 26 multiple-choice questions covering road legislation, car mechanics, traffic signs, and driving scenarios, with a pass threshold of 22 correct answers. Candidates typically prepare using question banks published by the GDDLR (General Directorate for Driving Licenses and Registrations) or online courses aligned with the Road Safety Legislation. However, many rely on repetitive memorization of answers without understanding the reasoning behind them, especially when questions involve legal terminology or complex visual elements such as traffic signs. This can lead to superficial learning and poor knowledge retention.

Although some online learning platforms provide limited explanations, these do not sufficiently cover all relevant legislation or justify the logic of correct answers. Additionally, no integrated and validated AI system exists to answer these driving exam questions reliably while explaining its reasoning. With recent progress in conversational AI and visual understanding, there is an opportunity to develop a multimodal system that retrieves relevant legal context and recognizes traffic signs automatically, offering contextualized answers with better accuracy.

This paper proposes and evaluates CarDrivingTutor - a multimodal framework that combines LLM-based question answering with semantic retrieval from the Romanian Road Safety Legislation and automatic detection of traffic signs using YOLOv8. Two models, Gemma3:27B and LLaVA:34B, are compared under three scenarios: (i) no context, (ii) text-based context, and (iii) text and visual context. Our research objective is to measure how contextual augmentation can improve the accuracy of answers on standardized Romanian driving license questions, providing a solid base for future educational tools that leverage AI.

To the best of our knowledge, this paper presents the first multimodal RAG-based framework that integrates semantic search in the Romanian Road Safety Legislation, visual traffic sign recognition using YOLOv8, and automated question answering with LLMs. This approach bridges the gap between rule-based preparation platforms and Intelligent Tutoring Systems.

RELATED WORK

Recent advances in Natural Language Processing have enabled large language models (LLMs) to perform complex reasoning, contextual understanding, and question answering across a wide range of domains. Systems such as GPT, LLaMA, or DeepSeek have achieved high performance in educational, legal, and medical applications when properly prompted. Retrieval-Augmented Generation (RAG) architectures [6] enhance this performance by combining an external retrieval mechanism with generative models. Instead of relying on memorized knowledge, the model is enriched with semantically retrieved context, which improves factual grounding and task-specific relevance. Recent surveys such as Cai et al. [1] and Corbière et al. [3] provide comprehensive overviews of RAG strategies across domains, highlighting the benefits of domain-adapted retrieval pipelines and reranking techniques in improving response quality, especially in context-heavy tasks like legal QA.

In the legal and educational domains, semantic retrieval has been used to assist users in understanding domain-specific texts. Gao et al. [4] introduced a system for legal QA based on semantic chunking of legislation, while Louis et al. [8] demonstrated improved performance on textbook-style MCQs by integrating textbook paragraphs as context via a vector search engine. Such systems have shown that context relevance is critical, especially when dealing with dense or technical language, such as that found in national traffic codes. Other studies like the one by Rekanar et al. [10] explored efficient legislative retrieval architectures for multilingual legal systems, further confirming the role of granular chunking and ranking in structured QA environments.

Multimodal approaches that combine textual and visual information have also shown promising results. Vision-language models like LLaVA [7] or Gemma3 [12] extend LLMs with visual encoders, allowing for image-grounded responses. These models have been applied in tasks like VQA (Visual Question Answering), captioning, and visual reasoning. For object detection, YOLOv8[1] is a state-of-the-art one-stage detector capable of real-time performance, widely used in traffic sign recognition, smart surveillance, and autonomous driving. Ibrahim and Kui [5] built a multi-sign detection framework based on YOLOv8 for real-world driving scenes, with over 240 sign classes.

In the context of driving education, prior work has focused mainly on user interfaces or simulator-based training, rather than automated QA. Most existing platforms in Romania rely on static multiple-choice databases, with some offering standard explanations for answers. However, these systems do not retrieve legal justification dynamically, nor do they process image-based questions with any form of visual reasoning.

METHOD

This section details the multimodal framework proposed for answering official driving exam questions in Romania. The system integrates textual context retrieval, visual sign detection, and LLM inference to generate accurate responses.

Dataset and Preprocessing

The dataset used for training, testing, and validation of the solution was automatically scraped from the platform of the Directorate of Driving License and Vehicle Registration (Romanian: Directia Regim Permise de Conducere si Inmatriculare a Vehiculelor – DRPCIV) school, from the category B, questions used in the real theoretical exam for obtaining a driving license[2].

All questions were saved in a MongoDB [2] collection imported locally in a standardized JSON format for an easier look at the experiment. Each question has the following fields: the question text, three answer options (A, B, C), the correct answer(s), a unique identifier provided by MongoDB, and the path to an associated image if it exists.

A total of 1,212 unique questions were collected, out of which 385 include an image, and the remaining 827 are only based on the text part. This dataset was saved in a MongoDB collection for a full overview of all questions explored in the project. Sample questions can be seen in Table 1. The dataset includes all possible questions that can be seen in the official driving license exam, making it a reliable benchmark for evaluation. It captures the capabilities of an LLM in a real-world educational domain that includes both textual and visual components.

Auxiliary Contextual Corpora

Two auxiliary corpora were also added to provide context for LLMs to answer questions using the retrieval-augmented generation method. The first one includes 671 articles from the Romanian Road Safety Legislation (Romanian: Codul Rutier)[3] and online courses explanations[4], which were extracted, parsed into individual paragraphs, so that each has its own individual meaning. Each article is stored as a separate document entry and has two fields: one is a string indicating the source and reference, and the other is the full legal text associated with the reference. This corpus is used for the textual RAG context. Table 2 shows a representation with a few articles as examples.

The second one stores 257 official definitions for all existing traffic signs[5]. The definitions are added in the context when YOLOv8 detects the corresponding sign. Traffic signs are stored with a name and official descriptions, as in Table 3.

Both corpora were processed and indexed using ChromaDB, a vector database optimized for semantic search, and saved

[1] https://huggingface.co/Ultralytics/YOLOv8

[2] https://scoaladrpciv.ro/intrebari/categoria-b

[3] https://www.drpciv-romania.ro/Code/Applications/web/index.cgi?action=codulrutier

[4] https://www.scoalarutiera.ro/curs-legislatie/

[5] https://www.codrutier.ro/semne-de-circulatie

Table 1. Sample Questions from the GDDLR Dataset.

ID	Question Text	Options	Correct	Image
Q1	When it snows heavily, what should you use?	A. position lights B. the horn C. the dipped-beam headlights	C	Yes
Q2	How should the driver proceed when driving on a road with three or more lanes in one direction?	A. drive on the lane next to the shoulder B. avoid driving on tram lines C. drive on the second or third lane at a speed lower than 100 km/h	A	No

Table 2. Sample Articles from Romanian Road Safety Legislation.

Article	Content
EGO - Article 63, para.(4)	Parking is considered the stopping of vehicles in specially arranged or designated and properly signaled spaces.
Regulation - Article 51, para.(1)	The green light allows passage.
Regulation - Article 52, para.(1)	The red light prohibits passage.

Table 3. Sample Traffic Sign Definitions.

Traffic Sign	Definition
Port Sign	Port Sign: Placed near a port.
Train Station Sign	Train Station Sign: Placed near a train station.

locally in JSON objects for better visualization. For embedding generation, the `all-MiniLM-L6-v2`[6] embedding model from SentenceTransformers was utilized, due to its high performance in retrieving semantically similar content. This ensured accurate matching between each question and the most relevant legal excerpt.

System Architecture

The system integrates several modules in a multimodal pipeline. For a better visualization of how the components interact with each other, Figure 1 presents a flowchart that illustrates the overall logic of the framework.

Retrieval-Augmented Generation

To support the automated generation process of LLMs, the framework integrates a Retrieval-Augmented Generation (RAG) pipeline designed to extract legal content semantically related to the question and answer choices. This mechanism ensures that relevant context is retrieved efficiently and dynamically, without relying on pre-written explanations or manually annotated data.

In the first stage, each input question is converted into a semantic vector using the `all-MiniLM-L6-v2` model. This vectorized representation includes both the question text and all associated answer options, which are concatenated into a single input string. The same embedding model is used to index the auxiliary corpus, consisting of 671 legislative paragraphs and 257 traffic sign definitions, preprocessed and stored in two ChromaDB collections: one for legal articles split at paragraph level, and another for traffic sign definitions indexed by name and official label.

Table 4. Framework Components Overview.

Component	Function
MongoDB	Stores all questions and model responses
YOLOv8	Detects traffic signs in input images
ChromaDB + MiniLM	Retrieves semantically similar law articles or sign definitions
CrossEncoder	Reranks documents based on relevance to the input question
Prompt Builder	Constructs input prompt for the LLM
LLM (Gemma/LLaVA)	Predicts answer using question + context

Once embedded, the question vector is compared to the indexed vectors using cosine similarity. The top 200 most similar paragraphs are retrieved as initial candidates for context. However, since cosine similarity alone may not provide optimal ranking in terms of contextual utility, the results are further refined through a reranking step.

In this reranking phase, a *cross-encoder/ms-marco-MiniLM-L6-v2* [11] model is used to re-evaluate each candidate by forming a pair between the question and the retrieved article. Each pair is then scored semantically. Documents with scores above 5.0 are considered highly relevant, and up to 10 of these are selected. If no documents exceed this threshold, the pipeline selects the top 3 documents (or fewer, if not available) that fall between 2.5 and 5.0. Any context scoring below 2.5 is discarded entirely. This selective thresholding helps maintain both specificity and brevity in the retrieved context. The final step involves concatenating the selected documents into a single text block, which becomes the RAG context fed into the LLM. If the input question also includes an image, the corresponding traffic sign definitions (extracted via YOLOv8 and matched semantically) are appended to this block, resulting in a full, enriched prompt. This dual-context

[6] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

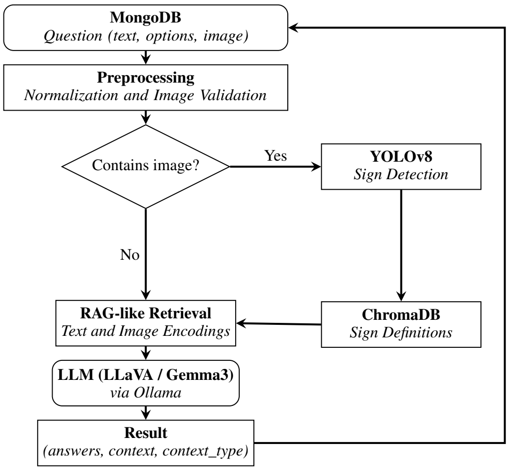


Figure 1. Multimodal question processing pipeline.

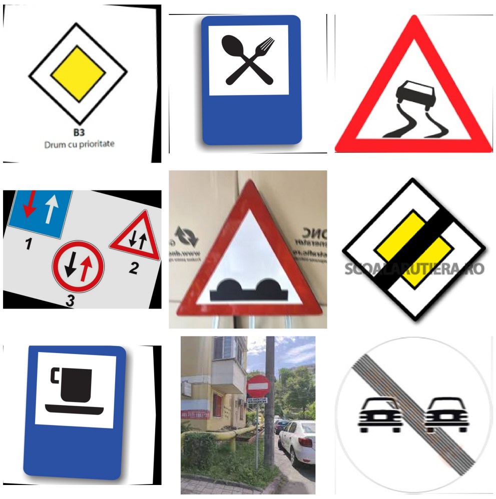


Figure 2. ***Sample images used for training YOLOv8 on Romanian traffic signs.***

design ensures that both textual and visual information are considered in the final LLM inference.

Visual Sign Detection

For questions that include images, the system integrates a visual recognition module that automatically identifies traffic signs present in the image. Detection is performed using YOLOv8, a one-stage object detection model known for its real-time inference speed and high accuracy. To improve its performance on real-world exam images, the model was fine-tuned on a custom dataset that includes traffic signs captured from various distances, angles, and lighting conditions. This diversity ensures that YOLOv8 can generalize well across the types of visuals commonly found in driving license questions.

The dataset was manually annotated using the Make Sense AI platform[7], where each traffic sign was labeled with its corresponding class name. The platform exports annotations in YOLO format using normalized bounding box coordinates, following the structure: `class_id` x_{center} y_{center} `width height`. These coordinates specify both the center and size of each object relative to the image, allowing precise localization. An example of the annotation and detection result is shown in Table 5.

A total of 5,719 images were used to train the YOLOv8 model, with an additional 365 images allocated for evaluation, following a 90%-10% split. The training was performed for 100 epochs on a dataset covering 242 unique traffic sign classes, corresponding to the Romanian road sign taxonomy. The output from YOLOv8 is later used in the prompt sent to the LLM by retrieving the official definitions of the detected signs from the auxiliary ChromaDB collection.

Prompt Construction and Instructions

After retrieving the relevant legal and/or visual context, the system builds a complete prompt adapted to the format of official multiple-choice questions. The prompt is composed of structured sections: the question text, answer options, the retrieved context (if available), and a set of clear instructions for the LLM to follow. This design proved to be the most efficient during testing and was adopted as the standard template throughout the evaluation.

The question and all three answer choices (A, B, C) are included in the prompt in a compact format. If context is available (either from the legal corpus or the sign definitions), it is added after the options. Then, concise instructions are appended, guiding the model to respond strictly with the correct option letters, in alphabetical order, without repeating or adding explanations.

Table 6 shows the two types of prompts used: one without context, and one with added textual/visual context. All instructions are formulated in Romanian, as this matches the language of the questions in the dataset.

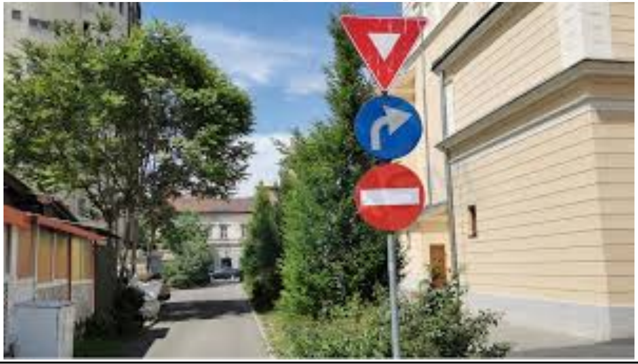
After iterative experiments, this became the standard prompt utilized in evaluation due to the best results. During the prompt construction stage, the study by Pezeshkpour and Hruschka [9] was also taken into account. This study shows that LLMs can be sensitive when it comes to multiple-choice questions and the order of the options. Their results show significant variations in the response accuracy, depending on the position of the correct answer, even in identical prompting conditions. In order to mitigate this effect, the order of the options is identical to the one on the GDDLR website, from which the questions were scraped.

Experimental Setup

Two multimodal models were tested: LLaVA:34B, fine-tuned for visual tasks, and Gemma3:27B, a newer model with instruction-tuned SigLIP visual encoder. Both support image input and are evaluated on three scenarios: no context, legal context only, and combined legal and visual context.

[7] https://www.makesense.ai

Table 5. YOLOv8 Labeling Output and Corresponding Detection Image



Class ID	Traffic Sign	x_{center}	y_{center}	Width	Height
43	Yield	0.599	0.131	0.161	0.249
0	Right Turn Only	0.607	0.552	0.132	0.193
112	No Entry	0.610	0.351	0.145	0.177

Table 6. Complete Prompt Instructions Used for LLM Inference (in English)

Prompt Type	Full Prompt Text
Without context	Question: {question['text']} Options: {formatted options} Instructions: Answer ONLY with the capital letters corresponding to the correct options, without explanations. Select only the correct options, in alphabetical order, without repeating letters. The answer format must be exactly: ["A"], ["B"], ["C"], ["A", "B"], ["A", "C"], ["B", "C"], ["A", "B", "C"].

With context	Question: {question['text']}
	Options: {formatted options}
	Context: {context}
	Instructions: Based on the context above, answer ONLY with the letters of the correct options, in alphabetical order.
	Select only the correct options, without repeating letters.
	The answer format must be exactly: ["A"], ["B"], ["C"], ["A", "B"], ["A", "C"], ["B", "C"], ["A", "B", "C"].

RESULTS

This research is meant to evaluate how well LLMs answer GDDLR questions without context and how much the context extracted with RAG and visual context detected by YOLOv8 in the sign definitions help. With all this taken into consideration, a discussion on which model performs better will be made. To evaluate the performance of LLMs in this multimodal framework, we will be looking at the results obtained after running both models (Gemma3:27B and LLaVA:34B). A representation of how many correct and incorrect answers each model has given can be seen in Table 7.

In order to better understand the contribution of the added context, a heatmap was constructed for the four scenarios and can be seen in Figure 3. The confusion matrices show how context impacted the performance of each model. Gemma3:27B benefited from adding textual context in a large number of cases for questions that did not include an image, as seen in the bottom left cell, more than LLaVA:34B. Furthermore, LLaVA failed to improve its results in questions with images.

Table 7. Correct versus Incorrect Answers With and Without Context

Model	Type	Correct	Wrong	Correct	Wrong
		(no)	(no)	(ctx)	(ctx)
Gemma3:27B	No image	426	401	532	295
LLaVA:34B	No image	380	447	423	404
Gemma3:27B	Image	177	208	196	189
LLaVA:34B	Image	181	204	180	205

Table 8 presents a comparative overview of the main performance metrics obtained by both models. It contains the percentages of correct answers with and without context, the proportion of cases where the context helped and harmed the answer, and a context reliability score calculated as a ratio of

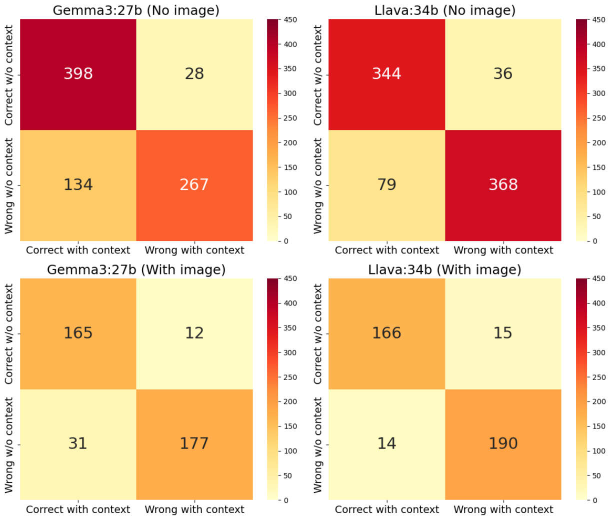


Figure 3. Heatmap showing the effect of added context on each model and question type.

cases where context either maintained or improved the answer, out of all cases where context could influence the outcome.

Experimental Optimizations

Going through the testing phase, adjustments were made to the prompting component and to the parameters used for the extracted context, in order to increase the general performance of the framework. On the prompting part, various variants were initially tested, including comprehensive and explicit explanations, but also summarized ones. The results on these prompts were inconsistent or generated invalid responses, like the same letter showing up twice in the response or the letter "D" appearing as a response even though the options are A, B, and C.

The final prompt is structured concisely to clearly emphasize the role of each component. This formula had the best results for the Gemma3 model when comparing the answers of the first 100 questions, and then was considered the standard in the evaluation and also in the comparison with LLaVA.

For the RAG component, the first 200 articles were selected based on cosine similarity between the semantic embedding of the question and the documents saved in ChromaDB. Smaller numbers, such as 50 or 100, were removing relevant articles that helped identify the correct answer to the analyzed questions, leading the LLM to provide an incorrect response. The top 200 provided sufficient coverage of relevant articles to be passed to CrossEncoder and to provide the best possible performance in the end. The thresholds used were chosen based on manual experiments to see what value a relevant score should have.

DISCUSSION

In this section, the performance for the two analyzed models, Gemma3:27B and LLaVA:34B, will be compared based on their capacity to correctly respond to the GDDLR questions in three different scenarios presented: without context, with textual context using RAG, and with context extracted from both visual and textual sources. Looking at the questions without context, Gemma3:27B model had 603 (426 without images, and 177 with images) correct answers, and LLaVA:34B had fewer, precisely 561 (380 without images and 181 with images) correct answers. This shows that Gemma3 is more capable of understanding the questions and knowing the answer without any auxiliary information.

Taking the questions with only the textual RAG given context, Gemma3 has 532 correct answers, marking 106 more than without context. LLaVA has 423 in this case, also increasing the number of correct answers by 43 compared to the previous scenario. These questions argue that Gemma3 utilizes the context from the Romanian Road Safety Legislation more effectively than its counterpart, which is less influenced by it. In the last question subset, where the images are considered and the detection of traffic signs is analyzed with YOLOv8, Gemma3 increases its number of correct answers by 19 from 177 to 196, and on the other hand, LLaVA loses one correct answer, going from 181 to 180. These results indicate that the road signs detection and their definitions introduced in the context don't significantly contribute to the final accuracy of the framework. Most likely, this phenomenon is caused by the fact that many pictures lack traffic signs that can be detected, and instead, they ask drivers to consider what they should do in complicated situations. There are also questions with images from a driver's perspective, and they do not involve road signs. Some examples of questions with pictures that confuse the LLMs can be seen in Table 9. The YOLOv8 model is not trained for this kind of question, and it is normal for it to give the wrong answer.

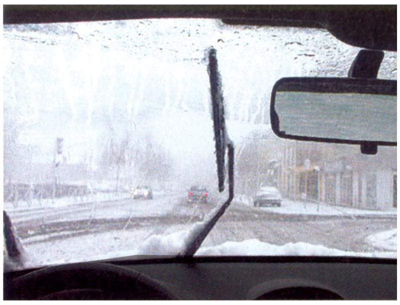
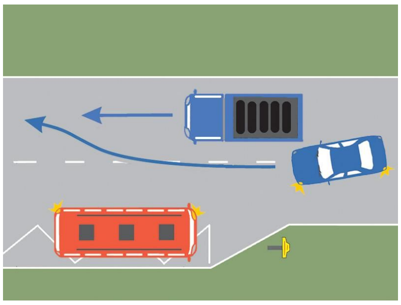
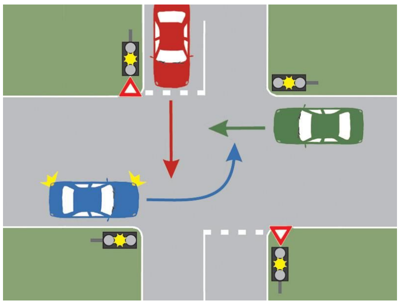
Limitations

Following the experimental evaluation, some limitations of the framework can be noted. One issue is context overload: in

Table 8. Model Evaluation Metrics Across Scenarios

Model	Acc. (ctx)	Acc. (no)	% Helped	% Harmed	Rel. Score
Gemma3:27B (no img)	0.64	0.52	16%	3%	0.95
LLaVA:34B (no img)	0.51	0.46	10%	4%	0.92
Gemma3:27B (img)	0.51	0.46	8%	3%	0.94
LLaVA:34B (img)	0.47	0.47	4%	4%	0.92

Table 9. Examples of Multiple-Choice Questions Used for Evaluation

Question 1	Question 2	Question 3
Text: When it snows heavily, what should you use? **A:** position lights **B:** the horn **C:** dipped-beam headlights **Gemma3:27b answer with context:** C **LLaVA:34b answer with context:** C **Correct:** C	**Text:** Is the car performing the overtaking maneuver correctly? **A:** no, because the bus has priority when leaving the station **B:** no, because this maneuver is not allowed at public transport stops **C:** yes, the maneuver is legal **Gemma3:27b answer with context:** B **LLaVA:34b answer with context:** A **Correct:** C	**Text:** In what order will the cars pass through the intersection when the traffic light shows flashing yellow? **A:** red, green, blue **B:** red, blue, green **C:** green, blue, red **Gemma3:27b answer with context:** A, B, C **LLaVA:34b answer with context:** A **Correct:** C
		

some cases, too long or imprecise context causes the model to generate wrong answers. This limitation was more often observed for the LLaVA model. Another challenge was visual context ambiguity. YOLOv8, while being a powerful model, was not optimally trained on this dataset. Given the 242 object classes, YOLOv8 would have learned more from more training images and longer training durations. In the current state, some misinterpretations and overlooks can be observed.

Further limitations involve prompting sensibility. LLMs have different responses based on the order in prompt components or on how the instructions are written, so it's not guaranteed that the optimal prompt for Gemma3 is also the best for LLaVA. Additionally, the Romanian traffic regulations do not provide enough information to answer the questions accurately. The documentation consists of articles and laws written in a juridic manner, difficult for making direct correlations with the questions that are written in a more colloquial language. Many questions are about practical situations or require car mechanics and signalization knowledge that is not in law articles and can be found in guides and teaching materials. All of this is affecting RAG efficiency because there are some cases in which there is not a single article relevant to the question and if the model doesn't know the answer without context, it won't know it with context either. Another possibility is to extract articles that are confusing the LLM when an article is over the threshold but has nothing to do with the question.

Finally, there are limitations from zero-shot inference. The models in this research have been used without any specific example or prior training for the GDDLR question format. They are answering only based on the given context from RAG and YOLOv8.

CONCLUSIONS AND FUTURE WORK

Current work proposed and evaluated CarDrivingTutor - an automated multimodal framework for answering GDDLR driving license exam questions by integrating natural language processing (NLP), computer vision (CV), contextual augmentation (RAG), and LLM inference. By comparing two large language models, Gemma3:27B and LLaVA:34B, the results proved that adding textual context significantly improves model performance. Gemma3 consistently outperformed LLaVA across all three tested scenarios: without context, with context retrieved through RAG, and with both visual and textual context.

Although the inclusion of visual context via YOLOv8 and sign definitions enhanced the model's ability to resolve some image-based questions, the most significant improvement came from the textual context, which provided meaningful legislative information that helped resolve ambiguities.

To conclude, the results attest to the viability of the proposed automated multimodal framework, while leaving clear room for future improvements, particularly in refining the contextual corpus, enhancing visual detection, and tailoring inference strategies through prompt engineering and fine-tuning.

The proposed framework could support further performance enhancements by considering possible improvements, such as fine-tuning the current model, experimenting with other visual detection models, or building a stronger knowledgebase. Fine-tuning on top of the GDDLR dataset could improve the learning of logical patterns. Introducing examples of questions and answers directly in the prompt, and classifying questions by type of context (e.g., Mechanical Knowledge, Traffic Sign Recognition, Traffic Signals, Traffic Legislation), can help build a more tailored context and guide the LLM toward better responses. Adding more explanatory content from official driving school materials to the article corpus could also broaden the scope of supported question types.

The visual detection module could be improved by employing more powerful object detection models, such as YOLOv8m, YOLOv8l, or DETR. These alternatives, trained on larger sets of labeled road sign images, could lead to more accurate sign recognition, particularly in real-world conditions or low-quality images. Lastly, extending the framework to also provide explanations for its answers could make it significantly more useful for educational purposes. Such explanations would help learners better understand their mistakes and consolidate knowledge in preparation for the driving exam.

ACKNOWLEDGMENTS

This research was supported by the project “Romanian Hub for Artificial Intelligence - HRIA”, Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 334906.

References

Tianhui Cai, Yifan Liu, Zewei Zhou, Haoxuan Ma, Seth Z Zhao, Zhiwen Wu, and Jiaqi Ma. 2024. Driving with regulation: Interpretable decision-making for autonomous vehicles with retrieval-augmented reasoning via llm. *arXiv preprint arXiv:2410.04759* (2024).

Anjali Chauhan. 2019. A review on various aspects of mongodb databases. *International Journal of Engineering Research & Technology (IJERT)* 8, 05 (2019), 90–92.

Charles Corbière, Simon Roburin, Syrielle Montariol, Antoine Bosselut, and Alexandre Alahi. 2025. Drivingvqa: Analyzing visual chain-of-thought reasoning of vision language models in real-world scenarios with driving theory tests. *arXiv e-prints* (2025), arXiv–2501.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997* 2, 1 (2023).

Baba Ibrahim and Zhou Kui. 2025. Enhancing Traffic Sign Recognition On The Performance Based On Yolov8. *arXiv preprint arXiv:2504.02884* (2025).

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, ¨ Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2024. Interpretable long-form legal question answering with retrieval-augmented large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 22266–22275.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483* (2023).

Kaavya Rekanar, Martin Hayes, Ganesh Sistu, and Ciaran Eising. 2024. Optimizing visual question answering models for driving: Bridging the gap between human and machine attention patterns. *arXiv preprint arXiv:2406.09203* (2024).

Jeanette Schofield, Shuyu Tian, Hoang Thanh Thanh Truong, and Maximilian Heil. 2025. DS@ GT at CheckThat! 2025: Exploring Retrieval and Reranking Pipelines for Scientific Claim Source Retrieval on Social Media Discourse. *arXiv preprint arXiv:2507.06563* (2025).

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786* (2025).

Jeromie Whalen, Chrystalla Mouza, and others. 2023. ChatGPT: Challenges, opportunities, and implications for teacher education. *Contemporary Issues in Technology and Teacher Education* 23, 1 (2023), 1–23.