RoD-TAL: A Benchmark for Answering Questions in Romanian Driving License Exams

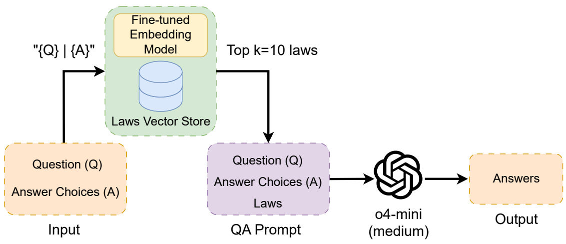
Anonymous ACL submission

Abstract

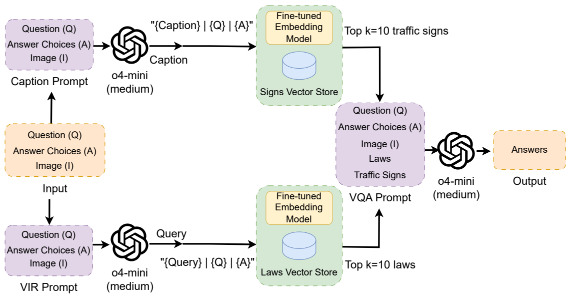
The intersection of AI and legal systems presents a growing need for tools that support legal education, particularly in under-resourced languages such as Romanian. In this work, we aim to evaluate the capabilities of Large Language Models (LLMs) and Vision-Language Models (VLMs) in understanding and reasoning about Romanian driving law through textual and visual question-answering tasks. To facilitate this, we introduce RoD-TAL, a novel multimodal dataset comprising Romanian driving test questions, text-based and image-based, alongside annotated legal references and human explanations. We implement and assess retrieval-augmented generation (RAG) pipelines, dense retrievers, and reasoning-optimized models across tasks including Information Retrieval (IR), Question Answering (QA), Visual IR, and Visual QA. Our experiments demonstrate that domain-specific fine-tuning significantly enhances retrieval performance. At the same time, chain-of-thought prompting and specialized reasoning models improve QA accuracy, surpassing the minimum grades required to pass driving exams. However, visual reasoning remains challenging, highlighting the potential and the limitations of applying LLMs and VLMs to legal education.

1 Introduction

The intersection of Artificial Intelligence (AI), legal systems, and Web technologies offers a powerful avenue for enhancing public access to structured legal knowledge. Road traffic law, in particular, provides a rule-based, codified domain that is well-suited for computational reasoning and the development of intelligent, web-based educational tools. As web information systems evolve to integrate data-driven AI models, legal education, especially in underserved linguistic and regional contexts, remains a vital yet underexplored application area (Lai et al., 2024).



(a) QA pipeline



(b) VQA pipeline

Figure 1: Architectures of the QA and VQA pipelines.

Despite significant advances in Large Language Models (LLMs) and retrieval-augmented architectures, their application to legally grounded question answering in low-resource languages remains limited (Hijazi et al., 2024; Das et al., 2024). In countries like Romania, where access to legal interpretation and educational resources is often inconsistent, there is a growing need for inclusive, intelligent systems that can support legal literacy and public understanding of the law (Guha et al., 2023; Hoppe et al., 2021).

Our research aims to evaluate LLMs and their reasoning capability on QA and VQA tasks, in the setting we propose, to assess how they can perform, their biases, their limitations, and how we can reliably integrate them to support education and law-related tasks.

We curated and created a novel dataset, called RoD-TAL, composed of **Ro**manian **D**riving **T**ests **a**nd **L**aws (annotated and referenced in the

Work/Author	Backbone	Modality	Accuracy				
			US	JP	CN	SG	RO
Zero-Shot QA (Zhou et al., 2024b)	GPT-4	Text	92.1%	86.5%	85.2%	88%	-
Zero-Shot VQA (Zhou et al., 2024a)	GPT-4V	Image	-	66.7%	79.2%	-	-
IDKB (Lu et al., 2025)	GPT-4o	Image	Overall for 15 languages: 53%				
RAG QA (Ours - Fig. 1a)	o4-mini	Text	-	-	-	-	86.4%
RAG VQA (Ours - Fig. 1b)	o4-mini	Image	-	-	-	-	78.3%

Table 1: Comparison of models and accuracy over languages. Results are reported for both text and image modalities, covering multiple languages (US - United States of America; JP - Japanese; CN - Chinese; SG - Singapore, RO - Romanian), with our method evaluated on Romanian for both modalities.

QA/VQA pairs), to help us assess our objectives.

The main contributions of this work are:

- We introduce a novel dataset called RoD-TAL;
- We evaluate LLMs in the context of low-resource language, Romanian, and a legal domain setting, and expose their biases and limitations;
- We propose a foundation for a legal and Romanian-based Dense Retriever which doesn't focus solely on semantic similarity but on Question and Legal Documents alignments;
- We propose solutions (see Figure 1 and Table 1) for all our identified Research Areas: IR, QA, VIR, VQA, with strong results in both IR and QA and less performant results on visual tasks.

2 Related Work

Driving QA and VQA. Zhou et al. (2024b) evaluated ChatGPT (GPT-4) (OpenAI et al., 2024b) on 814 written driving-license questions from California, Tokyo, Beijing, and Singapore, assessing performance across dimensions such as legal reasoning, situational understanding, and safety bias. Accuracy ranged from 85.2% to 92.1%, with lower scores on region-specific legal questions (e.g., 63.2% in China). While the model performed well overall, it showed limitations in handling local regulations and context-sensitive reasoning.

In follow-up work, Zhou et al. (2024a) assessed Vision-Language Models (VLMs) like ChatGPT and Bard on visual driving-license questions from Tokyo and Beijing. While models performed moderately on traffic sign recognition (70%) and better on scenario-based (80%) and combined visual tasks (80%), the study highlighted ongoing challenges in applying VLMs to real-world autonomous driving contexts.

However, these studies evaluate out-of-the-box LLM or VLM performance; they do not aim to optimize model accuracy and do not incorporate retrieval augmentation techniques. While GPT models have likely encountered many laws and question pairs during pretraining, reframing this task as a Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) problem could offer a more principled and scalable approach to improving legal and regulation-specific reasoning.

Vision-Language Driving Datasets. Multiple works (Kim et al., 2018; Deruyttere et al., 2019; Li et al., 2022; Malla et al., 2023; Qian et al., 2024; Sima et al., 2024; Shao et al., 2024; Tian et al., 2024; Park et al., 2024; Lu et al., 2025) are introducing and evaluating VL Driving Datasets for Autonomous Driving, some based on open QA, some based on MCQA, and spanning multiple languages, topics, and vehicle categories; see Table 2 for a comparison of these resources. Additionally, their limitation lies in not having references to legal corpora and relying solely on the internal knowledge of vision language models during QA tasks.

Multiple-Choice QA with LLMs Across Domains. Zhong et al. (2024) assessed LLMs, including GPT-4, on Chinese-based human exams like the Law School Admission Test (LSAT) and Lawyer Qualification Test (LQT), using formats such as multiple-choice and fill-in-the-blank. Among various prompting strategies, Few-Shot with Chain-of-Thought (CoT) performed best. GPT-4 scored 34–40% on the LQT and 31–87% on the LSAT. The study found stronger reasoning in high-resource languages (e.g., English) and domains like history and logic, while performance was weaker in law, math, and physics. Challenges included concept disambiguation, strict logical reasoning, and multi-hop inference, highlighting areas for further improvement. In the context of Romanian works, there has been a proposed MCQA dataset for the legal domain in Craciun et al. (2025) curated from legal examinations from different levels and spe-

Dataset	Data Type		Data Source		Data Domain			Knowledge Domain				Size	Law Ref.
	QA	MCQ	Real	Synthetic	Country	Lang.	Type	LR	SS	DT	DD		
BDD-X (Kim et al., 2018)	✓	✗	✓	✗	US	EN	Car	✗	✓	✗	✗	26K	✗
Talk2Car (Deruyttere et al., 2019)	✓	✗	✓	✗	US, SG	EN	Car	✗	✓	✗	✗	12K	✗
CODA-LM (Li et al., 2022)	✗	✓	✗	✓	DE, CN, SG	EN	Car	✗	✓	✗	✗	10K	✗
DRAMA (Malla et al., 2023)	✗	✗	✓	✗	JP	EN	Car	✗	✓	✗	✗	102K	✗
nuScenes-QA (Qian et al., 2024)	✓	✗	✓	✗	US, SG	EN	Car	✗	✓	✗	✗	460K	✗
DriveLM (Sima et al., 2024)	✓	✗	✓	✗	US, SG	EN	Car	✗	✓	✗	✗	2M	✗
LangAuto CARLA (Shao et al., 2024)	✓	✗	✗	✓	US	EN	Car	✗	✓	✗	✗	64K	✗
SUP-AD (Tian et al., 2024)	✗	✓	✓	✗	CN	EN	Car	✗	✓	✗	✗	-	✗
VLAAD (Park et al., 2024)	✗	✓	✓	✗	US	EN	Car	✗	✓	✗	✗	64K	✗
IDKB (Lu et al., 2025)	✓	✓	✓	✓	15	9	4	✓	✓	✓	✓	1M	✗
RoD-TAL (Ours)	✗	✓	✓	✗	RO	RO	Car	✓	✓	✓	✓	1.2K (400 w/ images)	✓

Table 2: Comparison of datasets, domains, and knowledge coverage, inspired by Lu et al. (2025). Knowledge Domain spans Laws and Regulation (LL), Signs and Signals (SS), Driving Techniques (DT), and Defense Driving (DD).

cializations, as well as an MCQA dataset for the medical domain (Dima et al., 2024), built from university entrance examinations.

Multiple-Choice VQA. Das et al. (2024) proposed EXAMS-V, a multilingual, multimodal benchmark for evaluating VLMs on multiple-choice questions across diverse domains (excluding law). GPT-4V scored an average of 42.5%, revealing limitations in multimodal reasoning and the integration of visual and textual information, despite showing early potential.

Information Retrieval and Retriever Fine-Tuning. Moreira et al. (2024) investigated fine-tuning dense retrievers for RAG tasks, emphasizing hard-negative mining strategies. Their results showed that starting from positive-aware setups and gradually introducing harder negatives significantly improved retrieval accuracy and response quality, underscoring the importance of training data curation in RAG pipelines.

Visual Information Retrieval. Dong et al. (2024) proposed a modality-aware retrieval approach that leverages visual LLMs to generate dense image captions for use in downstream querying, enhancing retrieval performance by integrating visual content more effectively.

3 RoD-TAL: Romanian Driving Tests And Laws

The RoD-TAL novel resource comprises a law corpus from the Romanian legislation, which we refer to as RoD-Law, and the QA part, including text- and image-based questions, referred to as RoD-QA.

3.1 RoD-Law Corpus

A central component of the RoD-TAL framework is its law-grounded foundation, the RoD-Law corpus, which ensures every answer can be explicitly traced back to the Romanian legislation. The legal corpus,

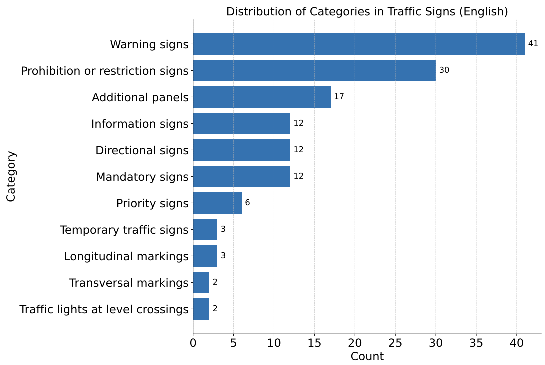


Figure 2: Distribution of traffic signs (indicators) per category.

curated as the retrieval base for all downstream tasks, was compiled from official sources, with abrogated sections removed to ensure relevance and currency, until March 2025. The composition of the legal corpus is summarized in Table 3, including laws from the traffic regulations, road code, penal code, and technical inspection law, and civil auto liability insurance law.

Corpus Source	Articles
Traffic Regulation Rules	225
Road Code	147
Penal Code (traffic-related)	9
Technical Inspection Law	15
Civil Auto Liability Insurance Law	47
Traffic Signs	140

Table 3: Distribution of articles in the RoD-Law corpus by legal source.

The RoD-Law legal corpus is complemented by an annotated collection of 140 distinct traffic signs, extracted from the answer references across the dataset and spanning 11 major categories (see Figure 2. Each sign is provided with its name, category, and a concise explanation, supporting both QA annotation and VQA tasks. Statistics are also presented in Appendix A

Dataset	Modality	Law Ref.	Size
Split 1	Text-based	✓	638
Split 2	Text-based	✗	131
Split 3	Image-based	✓	316
Split 4	Image-based	✗	71

Table 4: RoD-QA splits by modality and law reference annotation.

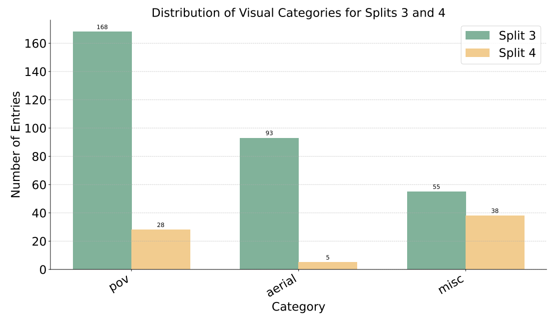


Figure 3: Number of visual questions per secondary category

3.2 RoD-QA Dataset

Dataset Structure. Built upon the curated legal base, the RoD-TAL dataset consists of multiple-choice questions sourced from the Romanian driving-license written tests, henceforth referred to as RoD-Law, available on the public educational platform **Scoala Rutiera**[1]. Each question is annotated with relevant legal references from RoD-Law, allowing for the evaluation of both standard LLM answering and retrieval-augmented generation (RAG) systems grounded in actual law.

The data structure for each sample includes the question, a set of candidate answers (with an explicit correct answer or more), explanation, legal reference, and a list of indicators where relevant. Visual questions were further categorized and enriched by o4-mini-based (OpenAI, 2025) sign annotation, followed by manual verification. For the experimental setup, we evaluate various combinations of text- and image-based questions with or without law annotations. An overview of the dataset splits is provided in Table 4.

This comprehensive annotation schema supports evaluation not only of LLM answer accuracy but also of legal retrieval precision/recall and legal grounding in RAG setups. The presence of visual (VQA) questions further enables research into legally grounded multimodal models, an emerging area in AI and law.

Dataset Statistics. The dataset encompasses 18 question categories in total, with the visual subset categorized into three additional secondary categories: point of view (pov), aerial, and miscellaneous (misc). Figure 3 details the distribution among visual-question secondary categories. Figure 7 from the Appendix A shows the number of questions per primary category. We further provide a fine-grained breakdown of questions by primary category and dataset split in Figure 4. Some categories are better represented than others depending on the split. For example, split 1 contains more questions related to *driver obligations* and *sanctions and offenses*, while the same categories contain fewer samples in the rest of the splits. For more statistics, refer to Appendix A.

4 Experiments

The problem of developing an AI system capable of answering legal questions based on Romanian traffic law can be decomposed into several distinct but interconnected tasks. These tasks span both textual and visual modalities, collectively defining the core components required for building, evaluating, and improving such a system. By segmenting the challenge into modular tasks, we facilitate targeted experimentation, fine-grained performance evaluation, and the possibility of optimizing each sub-component independently. We can pursue four topics: Information Retrieval (split 1), Question Answering (splits 1 and 2), Visual Information Retrieval (split 3), and Visual Question Answering (splits 3 and 4). We make our code publicly available[2].

4.1 Information Retrieval

An essential component of our system is the ability to retrieve relevant legal text passages that justify the correct answer to a question. Our focus is on optimizing Recall@k while minimizing k, since high values are impractical for downstream processing and the LLM generation part. Based on the distribution of legal articles that ground the questions (Figure 8 from Appendix A), we consider $k = 10$ sufficient for our experiments such that the LLM is context-bloated downstream.

In our work, we experiment with the embedding model $\text{mE5}_{\text{small}}$, where we evaluate different query building techniques, rerankers, or fine-tuning on

[1] https://www.scoalarutiera.ro/

[2] https://anonymous.4open.science/r/rodtal-exp-1F67

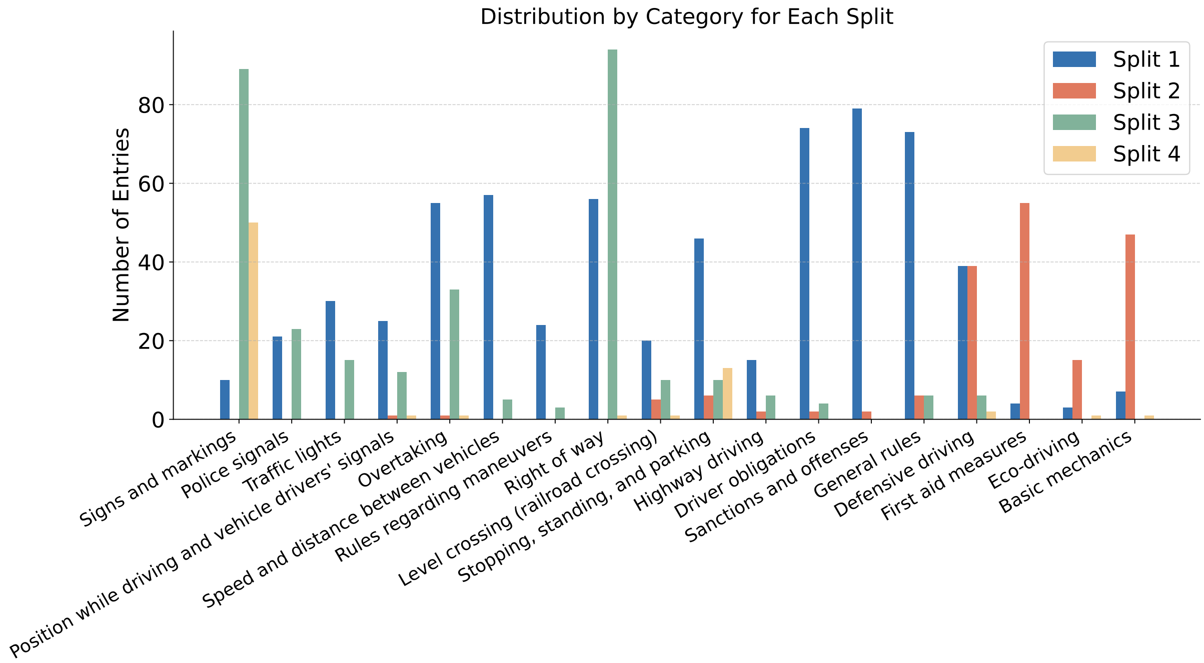


Figure 4: Number of questions per primary category and split.

real or augmented data. The full experimental details can be found in section B.4.

4.2 Question Answering

We utilize a CoT prompting strategy with the GPT-4o mini model, incorporating 10 retrieved documents alongside the question and answer options in the context. This setup reflects a standard RAG pipeline. To assess the value of retrieval, we experiment with a baseline prompt-based approach that does not employ document retrieval. This allows us to isolate the benefit of retrieval in performance. Additionally, we evaluate an *ideal RAG* setup, in which only the exact relevant documents are included, simulating ideal recall and precision conditions. We explore prompt engineering based on observed failure patterns in the model's behavior and propose a better prompt. We also conduct experiments without the CoT technique to test the role of step-by-step reasoning in answer accuracy. Lastly, we experiment with reasoning-tuned models under the same input specifications and prompts. We explore reasoning models (o4-mini) (OpenAI, 2025) with the same specifications and prompts as before to see if native reasoning helps with the problem. We also assess the performance of open-weights models, such as Mistral Small 3.1 Instruct[3] and Gemma 3 27B Instruct (Team et al., 2025), following the same scenarios as with the GPT-4o mini (OpenAI et al., 2024a) model.

[3] https://mistral.ai/news/mistral-small-3-1

4.3 Visual Information Retrieval

We adopt several methodologies to enhance a text query with the image characteristics. As a baseline, we compare against a plain QA search.

First, we generate and save 50-100-word captions using the o4-mini with CoT. We propose a combination of caption + QA. Then, we use the o4-mini model to rewrite our queries. We evaluate the following scenarios: Image + QA, Image + Caption + QA. For an extra measure, we also add the QA for the latter examples, in case the LLM omits some details

For testing traffic signs retrieval, a similar methodology is followed. However, we test two embedding scenarios: first, we only embed the traffic sign name and category, and second, we also append the explanation.

4.4 Visual Question Answering

We assess the performance of CoT prompting on the o4-mini model in three main configurations:

- **Model's prior knowledge**: The model answers questions using just the question, candidate answers, and image, without any external retrieval.
- **Ideal RAG**: The model is provided with exactly the relevant documents or legal references for each question (i.e., perfect retrieval

conditions).

- **Best RAG**: The model is provided with the best retrieval methodology from the VIR task for laws and traffic signs.

We also propose a similar methodology for the VIR task, comparing Caption + QA, Image + QA, and Image + Caption + QA, all leveraging CoT.

4.5 Evaluation Metrics

For information retrieval, we employ Recall@k, Precision@k, and nDCG@k (Järvelin and Kekäläinen, 2002). Recall@k measures the proportion of relevant documents among the top-k retrieved, while Precision@k evaluates the proportion of retrieved documents in the top-k that are relevant. nDCG@k further incorporates ranking position, assigning higher weight to relevant documents appearing earlier in the list. Out of these, the most important are recall@k and nDCG@k, while precision@k can be misleading due to the retrieval of more documents than needed.

For QA and VQA tasks, we use Precision, Recall, F1 score, and Exact Match (EM). Precision and Recall assess prediction accuracy and completeness, while F1 balances both. EM is our primary metric, requiring an exact match with the ground truth and offering no partial credit. Given our emphasis on strict answer correctness, EM is the central measure of QA performance.

4.6 Experimental Setup

Details about the experimental setup are presented in Appendix B. See Appendix C for full prompts in Romanian and translated to English. All the metric comments refer to the exact match score, which is also the performance that candidates are evaluated on in the MCQA setting.

5 Results

5.1 Information Retrieval

Our baseline experiment (1) in Table 5 yields only a 50% Recall@10 performance. Answer choices increase this performance by almost 10% in (2), thus a lot of the information needed for search is also present in the answer choices. Retrieving 40 documents and re-ranking the top 10, we also see a boost in performance by 8%-10%. Using an LLM to rephrase the query (5) yields similar results as with the QA (2); thus, the problem is semantic and not syntactic.

By fine-tuning our embedding model on the dataset, and now looking only at the test split, we see an 18% improvement for (6) in the Recall metric. Adding back a Reranker (7) hinders our performance by 11% (observed on split 2 specifically). This has its roots in the misalignment from which the Retriver model also suffered. The augmented fine-tuned model only yields 3% performance over its baseline (see (8) compared to (2)) but underperforms massively compared to the one trained on the data. This suggests a distribution shift from the original one: the complexity of the questions is lower, and the amount of related articles (as stated earlier) is much smaller.

These findings emphasize the importance of domain-specific representation learning for legal NLP in under-resourced languages and demonstrate that fine-tuning on even modestly sized, targeted datasets can significantly improve performance, more so than generic improvements in model architecture or reranking strategies.

5.2 Question Answering

Our first three experiments (1), (2), and 3 in Table 6 aim to compare the Retrieval component. On the first split, we observe that the RAG component improves performance by 11%, and comparing the best setup with an ideal RAG, the difference is marginal. This shows that even with not all documents retrieved or some over-retrieved and not relevant, the LLM is capable of completing the missing information and responding in a similar manner (only 2% variation).

Error analysis revealed three main failure modes (see also Appendix E):

- **Difficult Questions**: Difficulty handling nuanced or misleading options.
- **Safety Bias**: Prioritizing safe answers over legally correct ones.
- **Overthinking**: Overextending reasoning beyond the immediate question scope.

With this refinement in mind, using a better prompt to mitigate these findings, we achieve another 9% improvement over the previous results of strategy (4). At this step, we also want to ablate the CoT, making the model respond directly. This shows a loss of 25% in performance.

Adding a reasoning model over the previous experiments, we gain another 19% performance increase in strategy (6). Ablating Retrieval also shows it is still relevant, where we lose 22% in strategy (7).

Method	Retrieval - Train			Retrieval - Test		
	R@10	**P@10**	**nDCG@10**	**R@10**	**P@10**	**nDCG@10**
(1) Question based (Q)	50.15	11.45	38.75	49.29	11.01	37.48
(2) Question + Answer Choices (QA)	60.05	14.00	51.33	59.31	13.43	51.23
(3) QA + ReRanker jina	68.70	16.47	64.15	70.08	16.71	63.90
(4) QA + ReRanker bert-msmarco	63.99	15.21	56.22	65.60	15.54	56.49
(5) QA rephrased using GPT 4o-mini	60.84	14.27	51.70	60.64	14.14	49.86
(6) Finetuned Retriever	**99.89**	**27.21**	**99.71**	**88.14**	**23.28**	**81.41**
(7) Finetuned Retriever + ReRanker jina	80.43	20.25	71.44	77.55	19.92	70.23
(8) Augmented Finetuned Retriever	63.80	14.80	57.10	62.76	14.53	57.43

Table 5: Precision@10, Recall@10, and nDCG@10 metric scores for the tried and tested methods on the Information Retrieval task.

Method	Split 1		Split 2
	Train	**Test**	**Test**
(1) GPT-4o mini + CoT + RAG	57.8	55.5	72.9
(2) GPT-4o mini + CoT w/o RAG	46.1	43.0	71.3
(3) GPT-4o mini + CoT + Ideal RAG	59.0	63.3	69.6
(4) GPT-4o mini + CoT + RAG + better prompt	67.8	75.0	79.6
(5) GPT-4o mini + RAG + better prompt w/o CoT	42.7	47.7	60.8
(6) o4-mini + CoT + RAG + better prompt	**86.3**	**91.4**	**83.4**
(7) o4-mini + CoT + better prompt w/o RAG	64.3	63.3	82.3
(8) Mistral + CoT + RAG	42.5	42.2	51.9
(9) Mistral + CoT w/o RAG	46.1	39.8	68.0
(10) Mistral + CoT + Ideal RAG	30.4	35.2	4.4
(11) Mistral + CoT + RAG + better prompt	47.6	53.9	48.1
(12) Mistral + RAG + better prompt w/o CoT	57.5	53.9	75.7
(13) Gemma 3 + CoT + RAG	60.8	53.9	75.1
(14) Gemma 3 + CoT w/o RAG	48.0	38.3	65.7
(15) Gemma 3 + CoT + Ideal RAG	61.0	57.0	71.8
(16) Gemma 3 + CoT + RAG + better prompt	67.1	53.1	80.7
(17) Gemma 3 + RAG + better prompt w/o CoT	55.1	46.9	72.4

Table 6: Exact match score on IR/QA RAG pipeline.

Experiments with Mistral show similar patterns but overall weaker results. One thing to note is that this model would sometimes not follow the instruction, either by answering with the responses first and then arguing, or by starting an infinite loop of generating additional laws instead of answering. Also, for the experiments where we ablate CoT, the model would perform the same or better by 10-30%, depending on the dataset used.

Gemma achieves similar results compared to the OpenAI non-reasoning models.

5.3 Visual Information Retrieval

Using the fine-tuned retriever, our baseline text-based solution for retrieval yields a 60% recall, which is a strong initial result (strategy (1) in Table 7). Adding the caption to the search query improves the performance by 10%.

Further adding the image information, by using an LLM to rephrase the query (entries denoted with `R[...]` represent rephrasing of the query using VLM), shows an improvement in the Image + QA scenario by 3% in recall and a loss of 3% when also using the caption. Contrary to the belief that more context helps the model to produce a better query, it doesn't in this scenario.

Concatenating the QA pair to the previous experiments also shows a 7-8% improvement on both scenarios, suggesting that some details were left out in the rephrasing part.

Now, looking at retrieving traffic signs in Table 8, with the same experiments in mind, they follow a similar pattern in performance.

When changing the embedding method to also include more context and details for the indicators (the entries which have a *), we can see a slight improvement in the Caption + QA scenario. This combination helps with matching extra details in the caption to the extra details in the description, seeing a 3% gain in strategy (2*) over the previous best.

5.4 Visual Question Answering

We aim to evaluate VQA based on three prompting scenarios and three retrieval scenarios (Table 9). First, we want to test the input combination as before, either Caption+QA, Image+QA, or Image+Caption+QA. Consistently, Image+QA has

Method	Retrieval Laws - Split 3		
	R@10	P@10	nDCG@10
(1) Question + Answer Choices (QA)	60.45	12.72	56.18
(2) Caption + QA	70.30	15.06	56.41
(3) R[o4-mini + Image + QA]	73.51	16.13	58.38
(4) R[o4-mini + Image + Caption + QA]	67.70	14.87	52.85
(5) R[o4-mini + Image + QA] + QA	**77.09**	**16.83**	**62.67**
(6) R[o4-mini + Image + Caption + QA] + QA	75.17	16.32	59.67

Table 7: Precision@10, Recall@10 and nDCG@10 metric scores the Visual Information Retrieval of Laws task.

Method	Retrieval Indicators - Split 3			Retrieval Indicators - Split 4		
	R@10	P@10	nDCG@10	R@10	P@10	nDCG@10
(1) Question + Answer Choices (QA)	47.52	7.18	33.50	61.97	7.32	49.99
(2) Caption + QA	60.49	9.39	**46.62**	73.70	**9.29**	56.53
(3) R[o4-mini + Image + QA]	60.39	9.20	42.93	70.18	8.59	49.84
(4) R[o4-mini + Image + Caption + QA]	57.80	8.82	43.66	67.37	8.45	50.87
(5) R[o4-mini + Image + QA] + QA	60.23	9.24	46.14	73.47	9.01	56.29
(6) R[o4-mini + Image + Caption + QA] + QA	58.25	8.92	45.15	**74.41**	9.15	**60.38**
(1*) Question + Answer Choices (QA)	41.98	6.13	28.72	64.08	7.60	48.38
(2*) Caption + QA	**63.13**	**9.65**	44.96	69.95	8.59	53.13
(3*) R[o4-mini + Image + QA]	58.80	9.05	41.67	63.14	7.46	50.11
(4*) R[o4-mini + Image + Caption + QA]	57.25	8.95	39.95	68.54	8.16	50.63
(5*) R[o4-mini + Image + QA] + QA	58.93	8.98	41.93	70.89	8.59	54.89
(6*) R[o4-mini + Image + Caption + QA] + QA	59.34	9.11	42.13	72.06	8.59	56.11

Table 8: Precision@10, Recall@10 and nDCG@10 metric scores the Visual Information Retrieval of Indicators task.

a better performance than Caption+QA by 7-8%, while Image+Caption+QA is only 2-3% above Caption+QA. This is a similar pattern we spotted in the retrieval task, where more context (the caption) doesn't improve the performance, but it does affect performance.

Method	Split 3	Split 4
(1) o4-mini + Caption + QA + CoT	64.2	71.8
(2) o4-mini + Image + QA + CoT	71.5	74.6
(3) o4-mini + Image + Caption + QA + CoT	64.9	74.6
(4) o4-mini + Caption + QA + CoT + Ideal RAG	69.9	78.9
(5) o4-mini + Image + QA + CoT + Ideal RAG	**77.8**	78.9
(6) o4-mini + Image + Caption + QA + Ideal RAG	71.8	78.9
(7) o4-mini + Caption + QA + CoT + RAG	67.4	76.1
(8) o4-mini + Image + QA + CoT + RAG	75.6	**90.1**
(9) o4-mini + Image + Caption + QA + RAG	69.3	77.5

Table 9: Exact match scores on the VIR/VQA RAG pipeline using strategy (5) for laws and strategy (2*) for indicators

Checking on the Retrieval Component, we find the theoretical best and our best setup to be comparable in results (1-2% difference), while ablating RAG loses at least 5% performance.

5.5 Qualitative analysis

We present some samples of the QA examples where the better prompting strategy improved the results and mitigated the initially observed limitations in Appendix E. In the same Appendix, there are samples of wrong visual questions on each of the three secondary categories, using the best RAG setup.

6 Conclusions

In this work, we evaluateed Information Retrieval, Visual Information Retrieval, Question Answering, and Visual Question Answering both independently and in combination on a newly introduced dataset. Our results demonstrated promising performance across tasks, while also highlighting specific areas for improvement. In the IR component, further fine-tuning was needed to prevent the inclusion of actual positives during hard negative mining. For VIR, future work should explore more advanced methods incorporating joint image–text embeddings. In QA and VQA, improved prompting strategies or targeted model fine-tuning could help mitigate the limitations and biases inherent to current large language models.

Additionally, we examined the challenges LLMs face in multiple-choice QA settings, particularly the importance of minimizing over-selection to maximize precision. This research represents a novel contribution to the intersection of QA, VQA, IR, and VIR within the Romanian legal domain, specifically focused on traffic law, but with potential applicability to a wide range of legal tasks.

Limitations

We acknowledge that our work presents some limitations. The choice of not splitting larger documents into sub-chunks and truncating them instead aimed to simplify our experiments and benchmarking of the dataset, and we acknowledge that it might set us back from achieving a perfect retrieval result.

We also acknowledge that we didn't run the experiments with larger models, which could have improved the scores, due to hardware or budget constraints. For the same reasons, we didn't run the experiments on multiple seeds to explore variation.

Risks and Ethical Considerations

The only risk we see is in educational settings. It is known that LLMs suffer from hallucination (OpenAI et al., 2024a), for which our work doesn't fully address. If users were to learn for the driving exams using our setup, there is a chance of learning the wrong information, and this is also reflected in the results, which are perfect. In addition, our methods cannot be used for legal advice.

References

Cristian-George Craciun, Răzvan-Alexandru Smădu, Dumitru-Clementin Cercel, and Mihaela-Claudia Cercel. 2025. GRAF: Graph retrieval augmented by facts for Romanian legal multi-choice question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12708–12742, Vienna, Austria. Association for Computational Linguistics.

Rocktim Das, Simeon Hristov, Haonan Li, Dimitar Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. EXAMS-V: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. pages 7768–7791.

Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie-Francine Moens. 2019. Talk2Car: Taking control of your self-driving car. pages 2088–2098.

George-Andrei Dima, Andrei-Marius Avram, Cristian-George Craciun, and Dumitru-Clementin Cercel. 2024. RoQLlama: A lightweight Romanian adapted language model. pages 4531–4541.

Junnan Dong, Qinggang Zhang, Huachi Zhou, Daochen Zha, Pai Zheng, and Xiao Huang. 2024. Modality-aware integration with large language models for knowledge-based visual question answering. pages 2417–2429.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, and 1 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36:44123–44279.

Faris Hijazi, Somayah AlHarbi, Abdulaziz AlHussein, Harethah Abu Shairah, Reem Alzahrani, Hebah Al-Shamlan, George Turkiyyah, and Omar Knio. 2024. Arablegaleval: A multitask benchmark for assessing arabic legal knowledge in large language models. pages 225–249.

Christoph Hoppe, David Pelkmann, Nico Migenda, Daniel Hötte, and Wolfram Schenck. 2021. Towards intelligent legal advisors for document retrieval and question-answering in german legal documents. In *2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 29–32. IEEE.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. 2018. Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision (ECCV)*, pages 563–578.

Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S Yu. 2024. Large language models in law: A survey. *AI Open*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Kaican Li, Kai Chen, Haoyu Wang, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, and 1 others. 2022. Coda: A real-world road corner case dataset for object detection in autonomous driving. In *European conference on computer vision*, pages 406–423. Springer.

Yuhang Lu, Yichen Yao, Jiadong Tu, Jiangnan Shao, Yuexin Ma, and Xinge Zhu. 2025. Can lvlms obtain a driver's license? a benchmark towards reliable agi for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5838–5846.

Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. 2023. Drama: Joint risk localization and captioning in driving. In *Proceedings of* the IEEE/CVF winter conference on applications of computer vision, pages 1043–1052.

Gabriel de Souza P Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. 2024. Nv-retriever: Improving text embedding models with effective hard-negative mining. *arXiv preprint arXiv:2407.15831*.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. pages 2014–2037.

Rodrigo Nogueira and Kyunghyun Cho. 2020. Passage re-ranking with bert. *Preprint*, arXiv:1901.04085.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024a. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025. OpenAI o3 and o4-mini System Card. Technical report, OpenAI. System Card, released April 16, 2025.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

SungYeon Park, MinJae Lee, JiHyuk Kang, Hahyeon Choi, Yoonah Park, Juhwan Cho, Adam Lee, and DongKyu Kim. 2024. Vlaad: Vision and language assistant for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 980–987.

Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. 2024. Nuscenes-qa: A multimodal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4542–4550.

Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. 2024. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15120–15130.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. 2024. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pages 256–274. Springer.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. 2024. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. Agieval: A human-centric benchmark for evaluating foundation models, 2023. *URL https://arxiv. org/abs/2304.06364*.

Shanmin Zhou, Jialong Li, Takuto Yamauchi, Jinyu Cai, and Kenji Tei. 2024a. Evaluating vision-language models in visual comprehension for autonomous driving. In *2024 IEEE 4th International Conference on Software Engineering and Artificial Intelligence (SEAI)*, pages 205–209. IEEE.

Shanmin Zhou, Jialong Li, Mingyue Zhang, Daisuke Saito, Hironori Washizaki, and Kenji Tei. 2024b. Can chatgpt obey the traffic regulations? evaluating chatgpt's performance on driving-license written test. In *2024 International Conference on Innovation, Knowledge, and Management (ICIKM)*, pages 7–14. IEEE.

A Dataset Statistics

In this section, we present statistics of the RoD-TAL dataset, which comprises two main components: the legal corpus RoD-Law and the multiple-choice question-answering dataset RoD-QA. We analyze the characteristics and distributions relevant to each component to provide insights into the dataset's structure and content.

RoD-Law Statistics. The corpus totals 443 legal documents from the Romanian law. In Figure 5, we present the distribution of the number of tokens from the entire RoD-Law corpus. The distribution roughly follows a long-tail power law, where most documents have fewer tokens, with approximately 86% of documents containing text under 500 tokens in length.

Out of the 443 documents, 225 of them have references in the QA dataset. We present the token distribution in Figure 6. Similarly, 70% of the documents could enter a context window of 500 tokens.

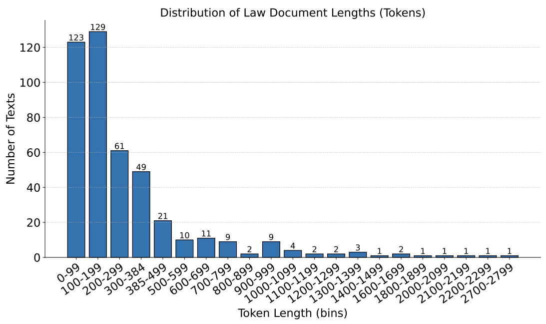


Figure 5: Number of tokens distribution of the entire RoD-Law corpus.

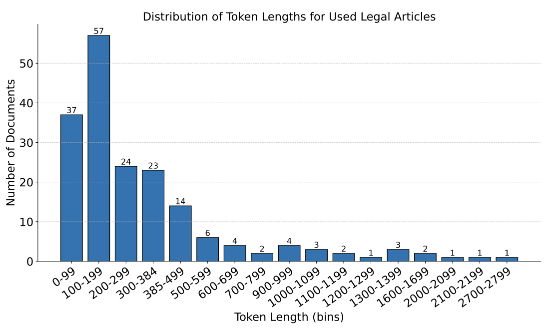


Figure 6: Number of tokens distribution on the documents referenced in RoD-QA.

RoD-QA Statistics. The QA dataset totals 1,156 samples containing text and images modalities with

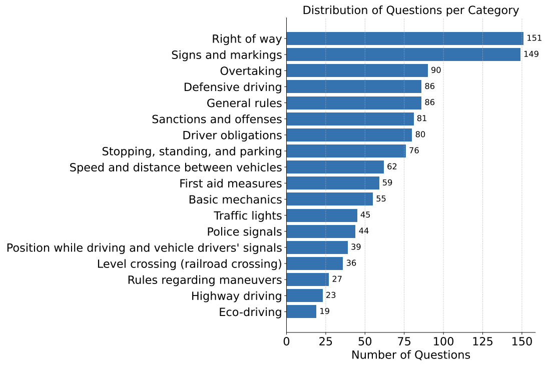


Figure 7: Number of questions per primary category

Figure 8: Distribution of legal article references per question

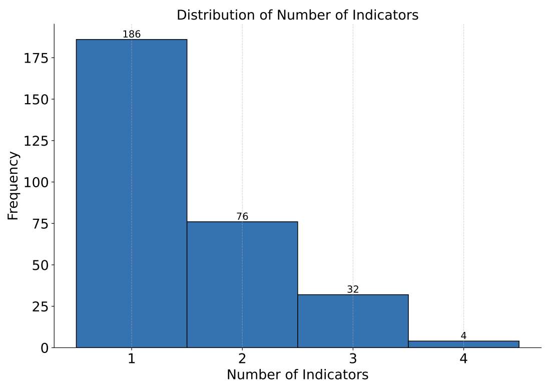


Figure 9: Distribution of traffic sign indicators per question

and without legal reference annotations, as presented in §3.2. In Figure 7, we showcase the distribution of questions per category. In total, there are 18 categories, most of them addressing *right of way* and *signs and markings*.

For the data annotated with legal references, we illustrate in Figure 8 the distribution per question. Most questions contain up to 10 references, which this also the main motivation to set $k = 10$ documents retrieved during experiments. In Figure 9,

Model	Num. Params.	Ctx. Size	Checkpoint
$mE5_{small}$	118M	512	multilingual-e5-small
Passage Reranking Multilingual BERT	168M	512	bert-multilingual-passage-reranking-msmarco
Jina Reranker v2	278M	1024	jina-reranker-v2-base-multilingual
Mistral Small 3.1	24B	128k	Mistral-Small-3.1-24B-Instruct-2503
Gemma 3 27B Instruct	27B	128k	gemma-3-27b-it
GPT-4o mini	undisclosed	128k	gpt-4o-mini-2024-07-18
o4-mini (medium)	undisclosed	200k	o4-mini-2025-04-16

Table 10: Model checkpoints used during experiments.

we present the distribution of the number of indicators illustrated in the questions containing images, where most questions show up to three indicators.

B Experimental Setup

B.1 Model Checkpoints

Table 10 presents the models and the checkpoints used during the experiments. We indicate the size as the number of parameters, context size in number of tokens, and the checkpoint on HuggingFace[4] or OpenAI Platform[5].

B.2 Hyperparameters

For the fine-tuning of the dense retriever model, we largely follow the settings from the original work, with several modifications specific to our setup. The model is trained using the Sentence Transformers library, with the following hyperparameters:

- **Number of epochs**: We train the first version of the model for 10 epochs, and the second one is early stopped after 1 epoch due to poor performance.
- **Batch size**: Both the training and evaluation batch sizes are set to 64.
- **Learning rate schedule**: We use a warmup ratio of 0.1.
- **Mixed precision**: Training is performed with FP16 precision enabled, while BF16 is disabled.
- **Batch sampler**: We use `BatchSamplers.NO_DUPLICATES` to avoid duplicate samples within a batch, which is beneficial for in-batch negative sampling losses.
- **Loss function**: The MultipleNegativesRankingLoss (InfoNCE) is used, with $mE5_{small}$ as the base model.
- **Evaluation and saving**: The model is evaluated and checkpoints are saved every 100 steps, with only the best two checkpoints retained.
- **Early stopping**: The best model is loaded at the end of training based on the evaluation metric (`eval_cosine_recall@10`), with `greater_is_better=True`.
- **Random seed**: To ensure reproducibility, all relevant random seeds (`torch`, `numpy`, `random`, and `transformers`) are set to 42, and deterministic training options are enabled.

[4] https://huggingface.co/
[5] https://platform.openai.com

B.3 Hardware Infrastructure and Computational Costs

We train the open-source LLM models on Google Colab with one NVIDIA A100 GPU and on 4 NVIDIA A100 GPUs from the institutional cluster, using the *vllm* package for LLM inference and better scalability.

For the closed-source LLM experiments, the GPT-4o mini model is used with a seed set to 25 and a temperature of 0. The o4-mini model is used without the possibility to set temperature or seed, as for now. The total cost for LLM experiments using the OpenAI API is $73.46.

B.4 Information Retrieval

Embedding model. We initially employ the $mE5_{small}$ multilingual model (Wang et al., 2024) to generate dense text embeddings for the legal corpus and user queries. This model supports Romanian and offers a sufficiently large context window to embed the relatively long articles typical of traffic law. It is also ranked among the top models on the MTEB leaderboard (Muennighoff et al., 2023) for retrieval tasks while supporting Romanian. We embed each article by concatenating the title metadata and content. We truncate the content if it is longer than the maximum context of 384, avoiding splitting into smaller chunks (as suggested in Figure 5).

Query formatting. To improve retrieval performance, we experiment with various input formulations, including concatenating the question with its answer options, which led to better embedding-based similarity scores.

Reranking. Further improvements were sought by incorporating multilingual reranking models: Jina Reranker v2[6] and Passage Reranking Multilingual BERT (Nogueira and Cho, 2020).

Query rewriting. We also explore query rewriting using LLMs, specifically GPT-4o mini via API (OpenAI et al., 2024a), attempting to rephrase questions in a way that aligns with the embedding model's representation space.

Fine-tuning retriever. Following poor results from these techniques, we hypothesize that domain and language mismatch are a core bottleneck, specifically in the specialized and underrepresented Romanian legal language. To address this, we fine-tune the $\text{mE5}_{\text{small}}$ model on our dataset for 10 epochs, 80-20 split, using the InfoNCE loss (Oord et al., 2018), consistent with the model's original training regime. We constructed a training set of approximately 6,960 samples, comprising **positive pairs** (i.e., questions with their correct legal references), and **hard negatives** (i.e., derived from top candidates retrieved by the base model, but judged incorrect – positive aware hard-negative mining, 5 each). We also test with reranking on top of the fine-tuned retriever.

Data augmentation. We further experiment with data augmentation via LLM-based synthesis. A few-shot prompt was created (2 examples of documents with associated questions and answers), and 1,000 sets of 2–6 legal documents were sampled. GPT-4o mini is used to generate 5 QA pairs per set. However, the LLM typically uses only 1–2 references per question, likely due to contextual incompatibility. After removing duplicates and entries with a similarity score over 0.98, we obtain 2,259 valid pairs, totaling 14,055 training samples. For this augmented dataset, we apply the same contrastive-based fine-tuning regime. The goal is not to leak any of the dataset distribution and statistics to our retriever and to validate based on the whole initial dataset.

[6]Jina Reranker V2

C Prompts (Romanian and English)

C.1 Information Retrieval Prompts

Romanian Version: VIR Rephrase query using LLM

Esti un politist rutier. Vorbesti doar Limba romana.
Primesti o grila de la un test auto, alaturi de raspunsurile posibile.
Scopul tau este sa generezi o singura intrebare astfel incat sa poti cauta legile care fac referinta la intrebarea primita.
Raspunsul tau se va incheia cu:
"Raspuns final: [intrebare generata tip string]"
Aceasta este intrebarea:
{question}
Aceastea sunt variantele de raspuns:
{answers}
===================

English Version: VIR Rephrase query using LLM

You are a traffic police officer. You only speak Romanian.
You receive a multiple-choice question from a driving test, along with the possible answers.
Your goal is to generate a single question that will help you search for the laws relevant to the received question.
Your answer must end with:
"Final answer: [generated question as string]"
This is the question:
{question}
These are the possible answers:
{answers}
===================

C.2 Question Answering Prompts

Romanian Version: QA Initial Prompt + RAG

Esti un politist rutier. Vorbesti doar Limba romana.
Trebuie sa rezolvi o grila de la un test auto. Aceasta grila poate avea unul sau mai multe raspunsuri corecte.
Gandestete care e raspunsul corect si raspunde la intrebare. La final, ultima parte din raspuns trebuie sa fie litera sau literele corecte.
De exemplu, raspunsul tau se va incheia cu
"Raspuns corect: A"
sau
"Raspuns corect: A,B"
Acesta este modul in care trebuie sa gandesti:
1. Citeste atent intrebare si variantele de raspuns.
2. Identifica ce informatii din legislatie ar putea fi relevante. (Legislatia Romaniei)
3. Daca ai mai multe raspunsuri corecte, argumenteaza fiecare alegere.
Aceasta este intrebarea:
{question}
Aceastea sunt variantele de raspuns:
{answers}
Aceastea sunt legile relevante, dar nu neaparat toate sunt relevante:
{documents}
===================

English Version: QA Initial Prompt + RAG

You are a traffic police officer. You only speak Romanian.
You need to solve a multiple-choice question from a driving test. This question may have one or more correct answers.
Think about what the correct answer is and respond to the question. At the end, the last part of your answer must be the correct letter or letters.
For example, your answer should end with
"Correct answer: A"
or
"Correct answer: A,B"
This is the way you should think:
1. Read the question and the answer options carefully.
2. Identify which information from the legislation might be relevant. (Romanian legislation)
3. If there are multiple correct answers, justify each choice.
This is the question:

{question}
These are the answer options:
{answers}
These are the relevant laws, but not all may be relevant:
{documents}
===================

Romanian Version: QA Initial Prompt without RAG

Esti un politist rutier. Vorbesti doar Limba romana.
Trebuie sa rezolvi o grila de la un test auto. Aceasta grila poate avea unul sau mai multe raspunsuri corecte.
Gandestete care e raspunsul corect si raspunde la intrebare. La final, ultima parte din raspuns trebuie sa fie litera sau literele corecte.
De exemplu, raspunsul tau se va incheia cu
"Raspuns corect: A"
sau
"Raspuns corect: A,B"
Acesta este modul in care trebuie sa gandesti:
1. Citeste atent intrebare si variantele de raspuns.
2. Identifica ce informatii din legislatie ar putea fi relevante. (Legislatia Romaniei)
3. Daca ai mai multe raspunsuri corecte, argumenteaza fiecare alegere.
Aceasta este intrebarea:
{question}
Aceastea sunt variantele de raspuns:
{answers}
===================

English Version: QA Initial Prompt without RAG

You are a traffic police officer. You only speak Romanian.
You need to solve a multiple-choice question from a driving test. This question may have one or more correct answers.
Think about what the correct answer is and respond to the question. At the end, the last part of your answer must be the correct letter or letters.
For example, your answer should end with
"Correct answer: A"
or
"Correct answer: A,B"
This is the way you should think:
1. Read the question and the answer options carefully.
2. Identify which information from the legislation might be relevant. (Romanian legislation)
3. If there are multiple correct answers, justify each choice.
This is the question:
{question}
These are the answer options:
{answers}
===================

Romanian Version: Enhanced QA prompt

Esti un politist rutier. Vorbesti doar in limba romana.
Trebuie sa rezolvi o grila de la un test auto. Grila poate avea unul sau mai multe raspunsuri corecte. Vei folosi strict legile din Romania.
Gandeste logic, dar nu extrapola peste informatiile oferite. Judeca doar momentul descris, nu presupune alte situatii.
Reguli de gandire:
1. Citeste cu maxima atentie intrebarea si variantele de raspuns.
2. Identifica strict ce prevederi din legislatia rutiera din Romania se aplica situatiei date.
3. Daca raspunsul pare "mai sigur" dar este contrar legislatiei, urmeaza legea, nu instinctul de precautie.
4. Alege DOAR raspunsurile care sunt complet corecte conform textului legii — nu ghici, nu completa informatii lipsa.
5. Daca un raspuns corect este mai bun decat altul dat ca si corect, include mai multe situatii specifice sau exceptii, atunci trebuie ales doar acela.
6. Argumenteaza clar de ce ai ales fiecare raspuns corect. Daca exista mai multe raspunsuri corecte, explica fiecare alegere separat.
7. Fii atent la mici detalii care pot schimba sensul intrebarii sau al raspunsurilor (exista intrebari-capcana).
La final, ultima parte din raspuns trebuie sa fie litera sau literele corecte.
De exemplu, raspunsul tau se va incheia cu:
"Raspuns corect: A"
sau
"Raspuns corect: A,B"
Aceasta este intrebarea:

{question}
Acestea sunt variantele de raspuns:
{answers}
Aceastea sunt legile relevante, dar nu neaparat toate sunt relevante:
{documents}
===================

English Version: Enhanced QA prompt

You are a traffic police officer. You only speak Romanian.
You need to solve a multiple-choice question from a driving test. The question may have one or more correct answers.
You will use only the laws from Romania.
Think logically, but do not extrapolate beyond the information provided. Judge only the described moment; do not assume other situations.
Thinking rules:
1. Read the question and answer choices very carefully.
2. Strictly identify which provisions of Romanian traffic legislation apply to the given situation.
3. If the answer seems "safer" but is contrary to the legislation, follow the law, not instinct.
4. Select ONLY the answers that are completely correct according to the letter of the law — do not guess, do not add missing information.
5. If one correct answer is better than another marked as correct, includes more specific situations or exceptions, then only that one should be chosen.
6. Clearly argue why you chose each correct answer. If there are multiple correct answers, explain each choice separately.
7. Pay attention to small details that can change the meaning of the question or answers (some questions are trick questions).
At the end, the last part of your answer must be the correct letter or letters.
For example, your answer should end with:
"Correct answer: A"
or
"Correct answer: A,B"
This is the question:
{question}
These are the answer choices:
{answers}
These are the relevant laws, but not all may be relevant:
{documents}
===================

Romanian Version: Better prompt without RAG

Esti un politist rutier. Vorbesti doar in limba romana.
Trebuie sa rezolvi o grila de la un test auto. Grila poate avea unul sau mai multe raspunsuri corecte. Vei folosi strict legile din Romania.
Gandeste logic, dar nu extrapola peste informatiile oferite. Judeca doar momentul descris, nu presupune alte situatii.
Reguli de gandire:
1. Citeste cu maxima atentie intrebarea si variantele de raspuns.
2. Identifica strict ce prevederi din legislatia rutiera din Romania se aplica situatiei date.
3. Daca raspunsul pare "mai sigur" dar este contrar legislatiei, urmeaza legea, nu instinctul de precautie.
4. Alege DOAR raspunsurile care sunt complet corecte conform textului legii — nu ghici, nu completa informatii lipsa.
5. Daca un raspuns corect este mai bun decat altul dat ca si corect, include mai multe situatii specifice sau exceptii, atunci trebuie ales doar acela.
6. Fii atent la mici detalii care pot schimba sensul intrebarii sau al raspunsurilor (exista intrebari-capcana).
Raspunde direct cu variantale corecte.
De exemplu, raspunsul tau se va incheia cu:
"Raspuns corect: A"
sau
"Raspuns corect: A,B"
Aceasta este intrebarea:
{question}
Acestea sunt variantele de raspuns:
{answers}
Aceastea sunt legile relevante, dar nu neaparat toate sunt relevante:
{documents}
===================

English Version: Better prompt without RAG

You are a traffic police officer. You only speak Romanian.
You need to solve a multiple-choice question from a driving test. The question may have one or more correct answers.
You will use only the laws from Romania.
Think logically, but do not extrapolate beyond the provided information. Judge only the described moment; do not assume other situations.
Thinking rules:
1. Read the question and answer options very carefully.
2. Strictly identify which provisions of Romanian traffic legislation apply to the given situation.
3. If the answer seems "safer" but is contrary to the legislation, follow the law, not instinct.
4. Select ONLY the answers that are completely correct according to the text of the law — do not guess, do not fill in missing information.
5. If one correct answer is better than another marked as correct, includes more specific situations or exceptions, then only that one should be chosen.
6. Pay attention to small details that can change the meaning of the question or answers (some questions are trick questions).
Respond directly with the correct options.
For example, your answer should end with:
"Correct answer: A"
or
"Correct answer: A,B"
This is the question:
{question}
These are the answer options:
{answers}
These are the relevant laws, but not all may be relevant:
{documents}
===================

C.3 Visual Information Retrieval Prompts

Romanian Version: Image Captioning

Descrie urmatoare imagine incluzand detalii legate de condus, indicatoare, ce se intampla in imagine.
Limiteaza-te la 50-100 cuvinte.
Foloseste numele indicatorului daca il cunosti, altfel o descriere succinta.

English Version: Image Captioning

Describe the following image, including details related to driving, road signs, and what is happening in the image.
Limit yourself to 50-100 words.
Use the name of the sign if you know it, otherwise provide a brief description.

Romanian Version: VIR by Image + QA

Esti un politist rutier. Vorbesti doar in limba romana.
Primesti o grila de la un test auto care are atasata si o imagine. Trebuie sa selectezi din imagine informatiile necesare, astfel incat sa imbuntatesti intrebarea originala si a facilita cautarea unor articole de lege relevante.
Include informatiile relevante despre situatie, indicatoare, si altele elemente specifice condusului si legii.
Reguli de gandire:
1. Citeste cu maxima atentie intrebarea si variantele de raspuns.
2. Analizeaza imaginea si extrage informatiile cele mai importante
3. Fii atent la mici detalii care pot schimba sensul intrebarii sau al raspunsurilor
4. Explica ce anume trebuie introdus in intrebare pentru a cauta informatiile corecte
La final, ultima parte din raspuns trebuie sa fie intrebarea reformulata.
"Raspuns final: [intrebare]"
Aceasta este intrebarea:
{question}
Acestea sunt variantele de raspuns:
{answers}
===================

English Version: VIR by Image + QA

You are a traffic police officer. You only speak Romanian.
You receive a multiple-choice question from a driving test, which also has an attached image. You need to select the necessary information from the image, so as to improve the original question and facilitate the search for relevant legal articles.

Include relevant information about the situation, road signs, and other elements specific to driving and the law.
Thinking rules:
1. Read the question and answer options very carefully.
2. Analyze the image and extract the most important information.
3. Pay attention to small details that may change the meaning of the question or the answers.
4. Explain what should be added to the question to help search for the correct information.
At the end, the last part of your answer must be the reformulated question.
"Final answer: [question]"
This is the question:
{question}
These are the answer options:
{answers}
===================

Romanian Version: VIR by Image + QA + Caption

Esti un politist rutier. Vorbesti doar in limba romana.
Primesti o grila de la un test auto care are atasata si o imagine. Trebuie sa selectezi din imagine informatiile necesare, astfel incat sa imbuntatesti intrebarea originala si a facilita cautarea unor articole de lege relevante.
Include informatiile relevante despre situatie, indicatoare, si altele elemente specifice condusului si legii.
Reguli de gandire:
1. Citeste cu maxima atentie intrebarea si variantele de raspuns.
2. Analizeaza imaginea si extrage informatiile cele mai importante
3. Fii atent la mici detalii care pot schimba sensul intrebarii sau al raspunsurilor
4. Explica ce anume trebuie introdus in intrebare pentru a cauta informatiile corecte
La final, ultima parte din raspuns trebuie sa fie intrebarea reformulata.
"Raspuns final: [intrebare]"
Aceasta este intrebarea:
{question}
Acestea sunt variantele de raspuns:
{answers}
Aceasta este descrierea imaginii:
{caption}
===================

English Version: VIR by Image + QA + Caption

You are a traffic police officer. You only speak Romanian.
You receive a multiple-choice question from a driving test, which also has an attached image. You need to select the necessary information from the image, so as to improve the original question and facilitate the search for relevant legal articles.
Include relevant information about the situation, road signs, and other elements specific to driving and the law.
Thinking rules:
1. Read the question and answer options very carefully.
2. Analyze the image and extract the most important information.
3. Pay attention to small details that may change the meaning of the question or the answers.
4. Explain what should be added to the question to help search for the correct information.
At the end, the last part of your answer must be the reformulated question.
"Final answer: [question]"
This is the question:
{question}
These are the answer options:
{answers}
This is the image caption:
{caption}
===================

C.4 Visual Question Answering Prompts

Romanian Version: Prompts for VQA without RAG

Esti un politist rutier. Vorbesti doar in limba romana.
Trebuie sa rezolvi o grila de la un test auto. Grila poate avea unul sau mai multe raspunsuri corecte. Vei folosi strict legile din Romania.
Vei primi o intrebare alaturi de o imagine, intrebarea avand stransa legatura cu intrebarea.
Gandeste logic, dar nu extrapola peste informatiile oferite. Judeca doar momentul descris, nu presupune alte situatii.
Reguli de gandire:
1. Citeste cu maxima atentie intrebarea si variantele de raspuns.
2. Identifica strict ce prevederi din legislatia rutiera din Romania se aplica situatiei date.

3. Daca raspunsul pare "mai sigur" dar este contrar legislatiei, urmeaza legea, nu instinctul de precautie.
4. Alege DOAR raspunsurile care sunt complet corecte conform textului legii — nu ghici, nu completa informatii lipsa.
5. Daca un raspuns corect este mai bun decat altul dat ca si corect, include mai multe situatii specifice sau exceptii, atunci trebuie ales doar acela.
6. Argumenteaza clar de ce ai ales fiecare raspuns corect. Daca exista mai multe raspunsuri corecte, explica fiecare alegere separat.
7. Fii atent la mici detalii care pot schimba sensul intrebarii sau al raspunsurilor (exista intrebari-capcana).
La final, ultima parte din raspuns trebuie sa fie litera sau literele corecte.
De exemplu, raspunsul tau se va incheia cu:
"Raspuns corect: A"
sau
"Raspuns corect: A,B"
Aceasta este intrebarea:
{question}
Acestea sunt variantele de raspuns:
{answers}
===================

English Version: Prompts for VQA without RAG

You are a traffic police officer. You only speak Romanian.
You need to solve a multiple-choice question from a driving test. The question may have one or more correct answers.
You will use only the laws from Romania.
You will receive a question along with an image; the question is closely related to the image.
Think logically, but do not extrapolate beyond the provided information. Judge only the described moment; do not assume other situations.
Thinking rules:
1. Read the question and answer options very carefully.
2. Strictly identify which provisions of Romanian traffic legislation apply to the given situation.
3. If the answer seems "safer" but is contrary to the legislation, follow the law, not instinct.
4. Select ONLY the answers that are completely correct according to the text of the law — do not guess, do not fill in missing information.
5. If one correct answer is better than another marked as correct, includes more specific situations or exceptions, then only that one should be chosen.
6. Clearly argue why you chose each correct answer. If there are multiple correct answers, explain each choice separately.
7. Pay attention to small details that can change the meaning of the question or answers (some questions are trick questions).
At the end, the last part of your answer must be the correct letter or letters.
For example, your answer should end with:
"Correct answer: A"
or
"Correct answer: A,B"
This is the question:
{question}
These are the answer options:
{answers}
===================

Romanian Version: Prompts for VQA with RAG

Esti un politist rutier. Vorbesti doar in limba romana.
Trebuie sa rezolvi o grila de la un test auto. Grila poate avea unul sau mai multe raspunsuri corecte. Vei folosi strict legile din Romania.
Vei primi o intrebare alaturi de o imagine, intrebarea avand stransa legatura cu intrebarea.
Gandeste logic, dar nu extrapola peste informatiile oferite. Judeca doar momentul descris, nu presupune alte situatii.
Reguli de gandire:
1. Citeste cu maxima atentie intrebarea si variantele de raspuns.
2. Identifica strict ce prevederi din legislatia rutiera din Romania se aplica situatiei date.
3. Daca raspunsul pare "mai sigur" dar este contrar legislatiei, urmeaza legea, nu instinctul de precautie.
4. Alege DOAR raspunsurile care sunt complet corecte conform textului legii — nu ghici, nu completa informatii lipsa.
5. Daca un raspuns corect este mai bun decat altul dat ca si corect, include mai multe situatii specifice sau exceptii, atunci trebuie ales doar acela.
6. Argumenteaza clar de ce ai ales fiecare raspuns corect. Daca exista mai multe raspunsuri corecte, explica fiecare alegere separat.
7. Fii atent la mici detalii care pot schimba sensul intrebarii sau al raspunsurilor (exista intrebari-capcana).
La final, ultima parte din raspuns trebuie sa fie litera sau literele corecte.
De exemplu, raspunsul tau se va incheia cu:
"Raspuns corect: A"

sau
"Raspuns corect: A,B"
Aceasta este intrebarea:
{question}
Acestea sunt variantele de raspuns:
{answers}
Aceastea sunt legile relevante, dar nu neaparat toate sunt relevante:
{documents_laws}
{documents_indicators}
===================

English Version: Prompts for VQA with RAG

You are a traffic police officer. You only speak Romanian.
You need to solve a multiple-choice question from a driving test. The question may have one or more correct answers.
You will use only the laws from Romania.
You will receive a question along with an image; the question is closely related to the image.
Think logically, but do not extrapolate beyond the provided information. Judge only the described moment; do not assume other situations.
Thinking rules:
1. Read the question and answer options very carefully.
2. Strictly identify which provisions of Romanian traffic legislation apply to the given situation.
3. If the answer seems "safer" but is contrary to the legislation, follow the law, not instinct.
4. Select ONLY the answers that are completely correct according to the text of the law — do not guess, do not fill in missing information.
5. If one correct answer is better than another marked as correct, includes more specific situations or exceptions, then only that one should be chosen.
6. Clearly argue why you chose each correct answer. If there are multiple correct answers, explain each choice separately.
7. Pay attention to small details that can change the meaning of the question or answers (some questions are trick questions).
At the end, the last part of your answer must be the correct letter or letters.
For example, your answer should end with:
"Correct answer: A"
or
"Correct answer: A,B"
This is the question:
{question}
These are the answer options:
{answers}
These are the relevant laws, but not all may be relevant:
{documents_laws}
{documents_indicators}
===================

D Additional Experimental Results

Method	Metric	Split 1		Split 2
		Train	**Test**	**Test**
(1) GPT-4o mini + CoT + RAG	Precision	77.0	74.7	85.5
(2) GPT-4o mini + CoT w/o RAG	Precision	69.0	67.3	85.2
(3) GPT-4o mini + CoT + Ideal RAG	Precision	76.5	76.8	82.7
(4) GPT-4o mini + CoT + RAG + better prompt	Precision	80.9	85.8	87.7
(5) GPT-4o mini + RAG + better prompt w/o CoT	Precision	68.5	70.2	79.2
(6) o4-mini + CoT + RAG + better prompt	Precision	**91.7**	**95.3**	**91.2**
(7) o4-mini + CoT + better prompt w/o RAG	Precision	76.8	77.1	89.2
(8) Mistral + CoT + RAG	Precision	65.2	66.8	70.2
(9) Mistral + CoT w/o RAG	Precision	80.3	78.6	91.0
(10) Mistral + CoT + Ideal RAG	Precision	46.9	51.7	18.5
(11) Mistral + CoT + RAG + better prompt	Precision	65.5	69.5	66.6
(12) Mistral + RAG + better prompt w/o CoT	Precision	88.6	86.8	93.4
(13) Gemma 3 + CoT + RAG	Precision	76.0	68.5	83.9
(14) Gemma 3 + CoT w/o RAG	Precision	63.0	58.4	75.5
(15) Gemma 3 + CoT + Ideal RAG	Precision	75.3	71.5	84.6
(16) Gemma 3 + CoT + RAG + better prompt	Precision	80.0	70.8	87.5
(17) Gemma 3 + RAG + better prompt w/o CoT	Precision	73.3	67.6	84.3
(1) GPT-4o mini + CoT + RAG	Recall	92.8	92.7	96.9
(2) GPT-4o mini + CoT w/o RAG	Recall	87.0	90.2	97.3
(3) GPT-4o mini + CoT + Ideal RAG	Recall	91.7	94.9	94.6
(4) GPT-4o mini + CoT + RAG + better prompt	Recall	90.7	92.8	94.6
(5) GPT-4o mini + RAG + better prompt w/o CoT	Recall	93.6	92.6	95.5
(6) o4-mini + CoT + RAG + better prompt	Recall	93.0	97.0	95.1
(7) o4-mini + CoT + better prompt w/o RAG	Recall	82.7	84.9	92.4
(8) Mistral + CoT + RAG	Recall	**97.6**	97.5	**98.6**
(9) Mistral + CoT w/o RAG	Recall	96.7	95.7	**98.6**
(10) Mistral + CoT + Ideal RAG	Recall	94.4	95.0	84.9
(11) Mistral + CoT + RAG + better prompt	Recall	97.1	**97.9**	97.2
(12) Mistral + RAG + better prompt w/o CoT	Recall	97.3	97.2	97.9
(13) Gemma 3 + CoT + RAG	Recall	89.4	87.1	93.9
(14) Gemma 3 + CoT w/o RAG	Recall	76.0	79.3	84.1
(15) Gemma 3 + CoT + Ideal RAG	Recall	88.8	88.9	94.6
(16) Gemma 3 + CoT + RAG + better prompt	Recall	90.5	87.2	90.7
(17) Gemma 3 + RAG + better prompt w/o CoT	Recall	88.8	88.3	92.0
(1) GPT-4o mini + CoT + RAG	F1-score	81.9	80.1	85.5
(2) GPT-4o mini + CoT w/o RAG	F1-score	74.1	74.1	89.1
(3) GPT-4o mini + CoT + Ideal RAG	F1-score	81.2	83.7	86.4
(4) GPT-4o mini + CoT + RAG + better prompt	F1-score	83.6	87.4	89.8
(5) GPT-4o mini + RAG + better prompt w/o CoT	F1-score	76.4	76.6	84.1
(6) o4-mini + CoT + RAG + better prompt	F1-score	**91.7**	**95.7**	92.2
(7) o4-mini + CoT + better prompt w/o RAG	F1-score	78.1	78.7	89.9
(8) Mistral + CoT + RAG	F1-score	69.1	70.9	73.6
(9) Mistral + CoT w/o RAG	F1-score	84.8	83.3	93.8
(10) Mistral + CoT + Ideal RAG	F1-score	50.8	55.2	23.4
(11) Mistral + CoT + RAG + better prompt	F1-score	68.2	72.4	69.6
(12) Mistral + RAG + better prompt w/o CoT	F1-score	92.0	90.7	**94.5**
(13) Gemma 3 + CoT + RAG	F1-score	79.7	73.8	86.7
(14) Gemma 3 + CoT w/o RAG	F1-score	66.5	64.2	77.9
(15) Gemma 3 + CoT + Ideal RAG	F1-score	79.2	76.3	87.5
(16) Gemma 3 + CoT + RAG + better prompt	F1-score	82.9	75.1	88.2
(17) Gemma 3 + RAG + better prompt w/o CoT	F1-score	77.8	73.3	86.3

Table 11: Precision, recall, and F1-score on the IR/QA RAG pipeline.

Method	Metric	Split 3	Split 4
(1) o4-mini + Caption + QA + CoT w/o RAG	Precision	65.8	72.5
(2) o4-mini + Image + QA + CoT w/o RAG	Precision	72.4	76.8
(3) o4-mini + Image + Caption + QA + CoT w/o RAG	Precision	66.3	74.6
(4) o4-mini + Caption + QA + CoT + Ideal RAG	Precision	70.7	80.3
(5) o4-mini + Image + QA + CoT + Ideal RAG	Precision	78.6	79.6
(6) o4-mini + Image + Caption + QA + Ideal RAG	Precision	73.1	78.9
(7) o4-mini + Caption + QA + CoT + RAG	Precision	68.5	76.1
(8) o4-mini + Image + QA + CoT + RAG	Precision	**76.3**	**90.1**
(9) o4-mini + Image + Caption + QA + RAG	Precision	69.8	78.2
(1) o4-mini + Caption + QA + CoT w/o RAG	Recall	67.4	73.2
(2) o4-mini + Image + QA + CoT w/o RAG	Recall	73.4	78.9
(3) o4-mini + Image + Caption + QA + CoT w/o RAG	Recall	67.7	74.6
(4) o4-mini + Caption + QA + CoT + Ideal RAG	Recall	71.5	81.7
(5) o4-mini + Image + QA + CoT + Ideal RAG	Recall	**79.4**	80.3
(6) o4-mini + Image + Caption + QA + RAG	Recall	74.4	78.9
(7) o4-mini + Caption + QA + CoT + RAG	Recall	69.6	76.1
(8) o4-mini + Image + QA + CoT + RAG	Recall	76.9	**90.1**
(9) o4-mini + Image + Caption + QA + RAG	Recall	70.3	78.9
(1) o4-mini + Caption + QA + CoT w/o RAG	F1-score	66.4	72.8
(2) o4-mini + Image + QA + CoT w/o RAG	F1-score	72.7	77.5
(3) o4-mini + Image + Caption + QA + CoT w/o RAG	F1-score	66.8	74.6
(4) o4-mini + Caption + QA + CoT + Idea RAG	F1-score	71.0	80.8
(5) o4-mini + Image + QA + CoT + Ideal RAG	F1-score	**78.9**	79.8
(6) o4-mini + Image + Caption + QA + Ideal RAG	F1-score	73.5	78.9
(7) o4-mini + Caption + QA + CoT + RAG	F1-score	68.9	76.1
(8) o4-mini + Image + QA + CoT + RAG	F1-score	76.5	**90.1**
(9) o4-mini + Image + Caption + QA + RAG	F1-score	69.9	78.4

Table 12: Precision, recall, and F1-score on the VIR/VQA RAG pipeline using strategy (5) for laws and strategy (2*) for indicators.

E Question and Answer Samples

This section highlights key examples of observed failure modes in strategy 1 (base LLM output) and their successful mitigation by strategy 4. Each example demonstrates a specific challenge in legal question-answering for LLMs and how a refined approach (strategy 4) can address these issues. Check the comments for every question for our observations.

Difficult Questions: These examples illustrate strategy 1's difficulty in discerning subtle nuances or misleading elements within questions or options, leading to incorrect or incomplete answers. Strategy 4 consistently provides more precise and legally accurate responses.

- **Question ID: d360db0b**
  - **Question:** How do you secure a car with a gear shifter so that it does not move on a slope?
  - **Strategy 1 Output (Translated):** To secure a car with a gear shifter so that it does not move on a slope, we must consider the safety measures provided by road legislation. 1. **Actuate the service brake** - This is a standard safety measure that helps keep the vehicle immobilized. The service brake must be used to prevent vehicle movement. 2. **Place the lever in "neutral"** - This is a measure that can help prevent vehicle movement, but it is not sufficient on its own, as a vehicle on a slope can still move if other safety measures are not also applied. 3. **Place the lever in a lower gear and actuate the parking or auxiliary brake** - This is a very effective measure, as using a lower gear helps keep the vehicle immobilized, and the parking or auxiliary brake provides additional safety. Considering the above, the correct answers are A and C, as both measures are essential to secure a car on a slope. Correct answer: A,C
  - **Strategy 4 Output (Translated):** To secure a car with a gear shifter so that it does not move on a slope, we must analyze each answer option in the context of Romanian road legislation. **A. Actuate the service brake;** - Actuating the service brake is a correct and necessary measure to stop the vehicle, but it is not sufficient to prevent its movement on a slope, especially if the vehicle is not properly immobilized. **B. Place the lever in "neutral";** - Placing the lever in "neutral" is not an effective measure to prevent vehicle movement on a slope. This does not ensure the immobilization of the vehicle, but only deactivates the transmission, which can lead to uncontrolled movement of the vehicle. **C. Place the lever in a lower gear and actuate the parking or auxiliary brake.** - This is the most correct option. According to legislation, in case of involuntary immobilization of the vehicle on a slope, the driver must engage a lower gear and actuate the auxiliary (or parking) brake. This ensures that the vehicle remains immobilized and will not move from the slope. **Argumentation:** - Answer A, although correct in the context of braking, is not sufficient to ensure complete immobilization of the vehicle on a slope. - Answer B does not comply with legal requirements for immobilizing a vehicle on a slope. - Answer C fully complies with road legislation, providing a complete and correct solution to prevent vehicle movement from a slope. Therefore, the correct answer is: Correct answer: C

- **Comments:** This case demonstrates Strategy 1's susceptibility to plausible but incomplete answers. While actuating the service brake is a safety measure, it is insufficient for legally securing a vehicle on a slope. Strategy 4 excels by providing a comprehensive and legally mandated solution, highlighting its improved precision in handling nuanced safety-related questions.

- **Question ID: 8be6e051**
  - **Question:** In which situations is overtaking prohibited?
  - **Strategy 1 Output (Translated):** To answer the question, we must analyze each answer option in the context of Romanian road legislation. A. **In uncontrolled intersections and at marked pedestrian crossings** - According to [Regulation-120], overtaking is prohibited in uncontrolled intersections and at marked pedestrian crossings. This is a clear rule prohibiting overtaking in these situations, so this option is correct. B. **In intersections and at a distance of 50m from them** - Also, according to [Regulation-120], overtaking is prohibited in intersections, but a distance of 50m from them is not explicitly mentioned. However, overtaking is prohibited near intersections, which may include a distance of 50m. This option is, therefore, correct. C. **In the action area of the "Speed Limit" sign** - This option is not correct, as the "Speed Limit" sign does not prohibit overtaking, but only limits the speed. Overtaking is not specifically prohibited in the action area of this sign, so this option is not correct. Therefore, the correct answers are A and B. Correct answer: A,B
  - **Strategy 4 Output (Translated):** To answer the question, we must analyze each answer option in the context of Romanian road legislation. **A. in uncontrolled intersections and at marked pedestrian crossings;** According to Art. 120 para. (1) lit. a) of the Regulation, overtaking is prohibited in uncontrolled intersections. Also, Art. 120 para. (1) lit. e) prohibits overtaking at marked pedestrian crossings. Thus, this option is correct. **B. in intersections and at a distance of 50m from them;** Art. 120 para. (1) mentions that overtaking is prohibited in intersections, but does not specify a distance of 50m from them. However, overtaking is prohibited near intersections, but no exact distance of 50m is specified. This option is partially correct, but not complete according to legislation. **C. in the action area of the "Speed Limit" sign.** The "Speed Limit" sign does not prohibit overtaking, but only limits the speed. Thus, overtaking is allowed in this area, respecting the speed limits. This option is incorrect. In conclusion, the only option that fully complies with Romanian road legislation is option A. Correct answer: A
  - **Comments:** This example illustrates Strategy 1 being misled by partially correct information. The 50m distance for overtaking prohibition around intersections, while intuitively plausible for safety, lacks specific legislative backing. Strategy 4's adherence to explicit legal text ensures accuracy and avoids overgeneralization of rules.

- **Question ID: 7753737d**
  - **Question:** In which situations is stopping prohibited?
  - **Strategy 1 Output (Translated):** To answer the question, we must analyze each answer option in the context of Romanian road legislation. A. **In the action area of the indicator that prohibits overtaking** - Stopping is prohibited in the action area of the "No Stopping" sign, but there is no explicit mention of stopping being prohibited in the action area of the indicator that prohibits overtaking. So, this option is not correct. B. **On one-way streets** - Stopping is not prohibited on one-way streets, as long as stopping and parking rules are respected. This option is not correct. C. **Next to another stopped vehicle, if this obstructs the circulation of two vehicles coming from opposite directions** - According to legislation, it is prohibited to stop next to another vehicle stopped on the roadway, if this obstructs the circulation of two vehicles coming from opposite directions. This option is correct. Therefore, the only correct option is C. Correct answer: C
  - **Strategy 4 Output (Translated):** To determine in which situations stopping is prohibited, we will analyze each answer option in accordance with Romanian road legislation. **A. in the action area of the indicator that prohibits overtaking;** According to Regulation 142, letter p), stopping is prohibited in places where overtaking is prohibited. This means that stopping in the action area of the indicator that prohibits overtaking is prohibited. So, this option is correct. **B. on one-way streets;** According to Regulation 141, para. (4), on one-way roads, voluntary stopping or parking of vehicles is permitted on the left side as well, as long as at least one lane of traffic remains free. Therefore, stopping on one-way streets is not prohibited, so this option is not correct. **C. next to another stopped vehicle, if this obstructs the circulation of two vehicles coming from opposite directions.** According to Regulation 142, letter h), it is not allowed to stop next to another vehicle stopped on the roadway, if this obstructs the circulation of two vehicles coming from opposite directions. This means that stopping in this situation is prohibited. So, this option is correct. In conclusion, the correct answers are A and C, as we have identified that stopping is prohibited in both situations mentioned. Correct answer: A,C
  - **Comments:** This case highlights a significant oversight by strategy 1, which failed to identify a legally correct situation for prohibited stopping (option A). Strategy 4's ability to correctly identify and justify this omission demonstrates its superior recall and application of specific legal regulations, proving critical in scenarios requiring comprehensive legal knowledge.

Safety Bias: These findings show strategy 1's tendency to prioritize answers based on general safety principles or common sense, even when not strictly aligned with specific legal provisions. Strategy 4 consistently corrects this by adhering to the precise legal framework.

- **Question ID: 4e0578a3**
  - **Question:** Penalty points are applied for:
  - **Strategy 1 Output (Translated):** For the question regarding the application of penalty points, we must analyze each answer option in the context of Romanian road legislation. A. **Disregarding the meaning of the "Children" sign** - This is a contravention that is sanctioned with penalty points, as it refers to the safety of children in traffic, and

disregarding this sign can lead to serious accidents. B. **Disregarding the meaning of the "Uneven Road" sign** - Although it is important to respect the meaning of this sign, its disregard is not specified in legislation as incurring penalty points. This is more of a recommendation for traffic safety, but it is not directly sanctioned with penalty points. C. **Disregarding the meaning of the "Stop" sign, installed at a level crossing** - This is a serious contravention, as disregarding this sign can lead to fatal accidents. According to legislation, disregarding the meaning of regulatory signals from railway agents directing traffic at level crossings is sanctioned with penalty points. Thus, the correct answers are A and C, as both refer to contraventions that incur penalty points according to road legislation. Correct answer: A,C

- **Strategy 4 Output (Translated):** For the question regarding penalty points, let's analyze each answer option in the context of Romanian road legislation. **A. disregarding the meaning of the "Children" sign;** - According to legislation, disregarding the meaning of the "Children" sign is not specified as a contravention that incurs penalty points. Although it is an important rule for children's safety, it is not explicitly mentioned in legislation that it would incur points. **B. disregarding the meaning of the "Uneven Road" sign;** - Similar to option A, disregarding the meaning of the "Uneven Road" sign is not mentioned in road legislation as a contravention that would incur penalty points. This is a warning for drivers, but it is not sanctioned with points. **C. disregarding the meaning of the "Stop" sign, installed at a level crossing.** - According to legislation, disregarding the meaning of regulatory signals from railway agents directing traffic at level crossings, including the "Stop" sign, is a contravention that incurs 6 penalty points. This is a strict rule, given the risks associated with level crossings. Therefore, the only correct option that aligns with Romanian road legislation is: Correct answer: C
- **Comments:** This is a salient example of Safety Bias. Strategy 1 incorrectly infers that disregarding a 'Children' sign should incur penalty points due to its clear safety implications, despite this not being explicitly stipulated in the legislation for penalty points. Strategy 4 accurately adheres to the precise legal text, demonstrating a critical distinction between general safety concerns and specific legal consequences.

- **Question ID: 713bbd86**
  - **Question:** While driving a car on a public road, it is forbidden:
  - **Strategy 1 Output (Translated):** For the question "While driving a car on a public road, it is forbidden:", let's analyze each answer option: A. **to transport children under 12 years old in the back seat** - This is not correct, because children under 12 years old must wear seat belts adapted to their weight and size, but it is not forbidden to transport them in the back seat, as long as they comply with regulations. B. **to smoke while driving** - This is correct. According to legislation, it is forbidden for the driver of a motor vehicle transporting dangerous goods or products to smoke while driving. Also, smoking can distract the driver's attention, which is dangerous. C. **front seat occupants to hold animals in their arms** - This is also correct. According to the regulation, it is forbidden for drivers of motor vehicles and persons occupying the front seat to hold animals in their arms while driving. Thus, the correct answers are B and C. Correct answer: B,C
  - **Strategy 4 Output (Translated):** To answer the question correctly, we must analyze each answer option in the context of Romanian road legislation. **A. To transport children under 12 years old in the back seat.** According to [Regulation-97], children under 12 years old must wear seat belts adapted to their weight and size, and those under 3 years old must be transported only in approved restraint systems. However, transporting children under 12 years old in the back seat is not prohibited, as long as these rules are respected. So, this option is not correct. **B. To smoke while driving.** According to [Regulation-176], it is forbidden for the driver of a motor vehicle transporting dangerous goods or products to smoke while driving. Although this rule applies specifically to the transport of dangerous goods, it is an unsafe and generally discouraged practice. However, it is not a general prohibition for all motor vehicles. This option is not correct. **C. Front seat occupants to hold animals in their arms.** According to [Regulation-97], it is forbidden for drivers of motor vehicles, as well as persons occupying the front seat, to hold animals in their arms while driving on public roads. This is a clear and direct prohibition, so this option is correct. In conclusion, the only correct option is C, as it refers to an explicit prohibition provided in road legislation. Correct answer: C
  - **Comments:** This case demonstrates strategy 1's Safety Bias by generalizing a specific legal prohibition (smoking for dangerous goods transporters) to all drivers, based on a general safety concern (distraction). Strategy 4's nuanced clarification highlights its precision in applying legal rules only where they are explicitly defined, preventing overreach based on perceived safety.

- **Question ID: d360db0b**
  - **Question:** How do you secure a car with a gear shifter so that it does not move on a slope?
  - **Strategy 1 Output (Translated):** To secure a car with a gear shifter so that it does not move on a slope, we must consider the safety measures provided by road legislation. 1. **Actuate the service brake** - This is a standard safety measure that helps keep the vehicle immobilized. The service brake must be used to prevent vehicle movement. 2. **Place the lever in "neutral"** - This is a measure that can help prevent vehicle movement, but it is not sufficient on its own, as a vehicle on a slope can still move if other safety measures are not also applied. 3. **Place the lever in a lower gear and actuate the parking or auxiliary brake** - This is a very effective measure, as using a lower gear helps keep the vehicle immobilized, and the parking or auxiliary brake provides additional safety. Considering the above, the correct answers are A and C, as both measures are essential to secure a car on a slope. Correct answer: A,C
  - **Strategy 4 Output (Translated):** To secure a car with a gear shifter so that it does not move on a slope, we must analyze each answer option in the context of Romanian road legislation. **A. Actuate the service brake;** - Actuating the service brake is a correct and necessary measure to stop the vehicle, but it is not sufficient to prevent its movement on a slope, especially if the vehicle is not properly immobilized. **B. Place the lever in "neutral";** - Placing the lever in "neutral" is not an effective measure to prevent vehicle movement on a slope. This does not ensure the immobilization of the vehicle, but only deactivates the transmission, which can lead to uncontrolled movement of the vehicle. **C. Place the lever in a lower gear and actuate the parking or auxiliary brake.** - This is the most correct option. According to legislation, in case of involuntary immobilization of the vehicle on a slope, the driver must engage a lower gear and

actuate the auxiliary (or parking) brake. This ensures that the vehicle remains immobilized and will not move from the slope. **Argumentation:** - Answer A, although correct in the context of braking, is not sufficient to ensure complete immobilization of the vehicle on a slope. - Answer B does not comply with legal requirements for immobilizing a vehicle on a slope. - Answer C fully complies with road legislation, offering a complete and correct solution to prevent vehicle movement from a slope. Therefore, the correct answer is: Correct answer: C

- **Comments:** This example, also categorized under "Tricky Questions", demonstrates strategy 1's Safety Bias by accepting "actuating the service brake" as a sufficient safety measure for vehicle immobilization on a slope. While seemingly safe, it is legally incomplete for long-term securement. Strategy 4, by providing the comprehensive legal requirement (engaging a lower gear and parking brake), showcases its rigorous adherence to mandated procedures over merely plausible safety practices.

Overthinking: These examples highlight instances where strategy 1 introduces extraneous information, overgeneralizes specific rules, or misinterprets the scope of a question, leading to less precise or incorrect answers. Strategy 4 provides concise and legally accurate responses.

- **Question ID: b596b521**
  - **Question:** What obligations does the driver have when driving on a public road?
  - **Strategy 1 Output (Translated):** To answer the question, we must analyze each answer option in the context of Romanian road legislation. A. "to drive only if the monthly medical check-up is performed;" - This statement is not correct. Drivers are required to undergo periodic medical examinations, but there is no obligation to undergo a monthly medical check-up. Medical examinations are set at longer intervals, usually annually or every 5 years, depending on age and type of license. B. "to drive only on road sections where access is permitted and to comply with regulations on maximum authorized total masses admitted by the competent authority;" - This statement is correct. According to legislation, drivers must respect regulations regarding access to certain road sections and comply with norms regarding maximum authorized total masses. C. "to inquire in advance, from road administrators, about any maximum and minimum speed limits." - This statement is also correct. Drivers have an obligation to be informed about speed limits, which are set by road administrators. Therefore, the correct answers are B and C. Correct answer: B,C
  - **Strategy 4 Output (Translated):** To answer the question correctly, we must analyze each answer option in the context of Romanian road legislation. **A. to drive only if the monthly medical check-up is performed;** - This statement is not correct. According to legislation, drivers must undergo periodic medical examinations, but there is no obligation to undergo a monthly medical check-up. Medical examinations are set at longer intervals, usually annually or at the request of competent authorities, depending on the vehicle category. Therefore, this option does not comply with the legislation. **B. to drive only on road sections where access is permitted and to comply with regulations on maximum authorized total masses admitted by the competent authority;** - This statement is correct. According to Regulation-147, the driver of a motor vehicle is obliged to drive only on road sections where access is permitted and to comply with regulations on maximum authorized total masses. This is a clear obligation stipulated in legislation. **C. to inquire in advance, from road administrators, about any maximum and minimum speed limits.** - This statement is partially correct, but it is not an explicit obligation of the driver. Although it is recommended to inquire about speed limits, the legislation does not explicitly impose this obligation. Drivers are responsible for respecting traffic signs and road indicators, which indicate speed limits, but there is no formal obligation to inquire in advance from the road administrators. In conclusion, the only correct option according to Romanian road legislation is: **Correct answer: B**
  - **Comments:** Strategy 1 exhibits Overthinking by elevating a recommended practice (informing oneself about speed limits) to a formal legal obligation. This illustrates a tendency to extend common-sense guidance into codified law. Strategy 4 correctly distinguishes between general good practice and explicit legal requirements, showing improved discernment.

- **Question ID: 33615649**
  - **Question:** Changing direction to the left, in an intersection not provided with guiding markings, is done:
  - **Strategy 1 Output (Translated):** To answer the question correctly, we must analyze each answer option in the context of Romanian road legislation. 1. **A**: "by circumventing the imaginary center of the intersection" - This option is correct, because changing direction to the left is done by circumventing the imaginary center of the intersection, according to Regulation 110, which stipulates that changing direction to the left is done through the left of the imaginary center of the intersection. 2. **B**: "through the left of the imaginary center of the intersection, without intersecting the trajectory of those moving from the opposite direction and turning left" - This option is also correct. According to the same regulation, it is important that the maneuver is performed without intersecting the trajectory of vehicles coming from the opposite direction. 3. **C**: "by free agreement with those moving from the opposite direction" - This option is not correct. Changing direction cannot be done by free agreement, but traffic rules and the meaning of markings and indicators must be respected. Therefore, the correct answers are A and B. Correct answer: A,B
  - **Strategy 4 Output (Translated):** For the question regarding changing direction to the left in an intersection not provided with guiding markings, let's analyze each answer option in the context of Romanian road legislation. **A. by circumventing the imaginary center of the intersection;** This option is not correct. According to Regulation 110, changing direction to the left is done through the left of the imaginary center of the intersection, but circumvention is not mentioned. Circumvention could imply a maneuver that does not comply with traffic rules. **B. through the left of the imaginary center of the intersection, without intersecting the trajectory of those moving from the opposite direction and turning left;** This option is correct. According to Regulation 110, changing direction to the left is performed through the left of the center of the intersection, without intersecting the trajectory of vehicles coming from the opposite direction and turning left. This is an essential rule to avoid collisions. **C. by free agreement with those moving from the opposite direction.** This option is not correct. Changing direction cannot be done "by free agreement," as road

legislation imposes strict rules to ensure traffic safety. Each driver must respect traffic rules and not rely on informal agreements with other road users. In conclusion, the only correct option according to Romanian road legislation is: Correct answer: B

  - **Comments:** This case demonstrates strategy 1's Overthinking by misinterpreting a specific legal maneuver. Strategy 1 incorrectly validates "circumventing the imaginary center of the intersection," whereas strategy 4 correctly clarifies that the maneuver is simply "through the left of the imaginary center." This indicates a tendency to over-complicate or misapply procedural details from the legal text.

- **Question ID: 4fb7ec2c**
  - **Question:** The yellow traffic light signal:
  - **Strategy 1 Output (Translated):** For the question regarding the yellow traffic light signal, let's analyze each answer option: A. "Permits entry into the intersection, if it follows the green light." - This is correct, because the yellow signal appears after green and indicates that the vehicle can prepare to stop, but it does not prohibit entry into the intersection if the vehicle is too close to stop safely. B. "Permits changing direction to the right." - This is not a generally correct statement, because the yellow signal does not explicitly permit changing direction. Changing direction must be done in compliance with traffic rules and signaling. C. "Prohibits entry into the intersection, if the vehicle can be stopped safely." - This is correct, according to legislation, the yellow signal indicates that the vehicle must stop, unless it cannot do so safely. Thus, the correct answers are A and C. Correct answer: A,C
  - **Strategy 4 Output (Translated):** For the question regarding the yellow traffic light signal, let's analyze each answer option in the context of Romanian road legislation. **A) permits entry into the intersection, if it follows the green light;** - This statement is not correct. The yellow signal does not permit entry into the intersection, but rather informs the driver that they must prepare to stop, unless they are so close to the intersection that they could not stop safely. Therefore, it cannot be considered that it permits entry into the intersection. **B) permits changing direction to the right;** - This statement is partially correct, but it is not complete. The yellow signal does not automatically permit changing direction, but rather indicates that the vehicle must prepare to stop. The driver can change direction, but must comply with traffic rules and ensure the maneuver is safe. **C) prohibits entry into the intersection, if the vehicle can be stopped safely.** - This statement is correct. According to legislation, the yellow signal prohibits entry into the intersection if the vehicle can be stopped safely. This is a clear rule stipul...
  - **Comments:** Strategy 1's interpretation of the yellow traffic light signal exemplifies Overthinking, specifically by over-permitting action. It incorrectly states that the yellow light "permits entry into the intersection" if it follows a green light, attempting to rationalize a scenario where stopping safely isn't possible. Strategy 4 precisely clarifies that the yellow light's primary function is to "prohibit entry if safe to stop," demonstrating a critical adherence to legal prohibitions over conditional permissions.

Image-based QA: These examples highlight instances where images were part of the question.

- **Question ID: b3c12b4c**
  - **Category: aerial**
  - **Image:** Figure 10
  - **Question:** Which of the vehicles may continue through the intersection?
    A: The trams, because they have right of way;
    B: The blue car;
    C: The blue car and the yellow car.

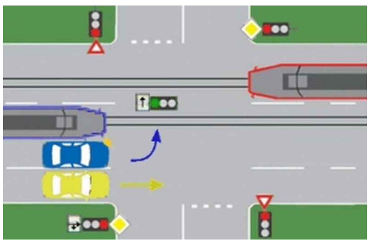


Figure 10: Image corresponding to question ID: b3c12b4c

  - **Reasoning:** [REASONING]**Analyzing tram and car directions**
    I'm examining the red and purple trams, which are perpendicular to cars. The rails run east-west, and the trams go horizontally in that direction. The blue car on the road is signaling with its right indicator blinking. The arrow near it curves left, indicating it wants to turn left. However, it seems the blue car is actually signaling a right turn as it approaches the intersection. The yellow car is going straight, intending to turn right at the intersection. [REASONING]**Interpreting traffic signals and car directions**
    In the image, the curved arrow from the blue car indicates it's turning right towards the tram tracks. The yellow car has a straight arrow and is going straight through the intersection. The blue car is heading east, while the yellow car continues north. The purple tram approaches from the left, going east, and the red tram comes from the east, heading west. There are traffic lights showing green for the trams, allowing them to proceed straight, while signals for cars entering from the east are less clear. [REASONING]**Deciphering tram signals and traffic priorities**
    I'm evaluating the tram signals. On the right, the light has a yellow diamond sign indicating priority but also shows a green circle for the red tram. So, the tram can proceed. The main road runs east-west, confirming trams travel on

it. The trams (red and purple) have priority with the green light, allowing them to move without conflict. Cars on the north-south approach must stop due to the red signal, and the blue car appears to yield before turning onto the main road.
  - **Output:** Drivers of the cars are stopped by the traffic light (red color) and, in addition, are on the non-priority road (indicated by the "yield" arrow); the trams, on the other hand, have a green signal at their dedicated traffic light and travel on the priority road (marked with the "priority road" sign). The intersection is controlled by traffic lights, so the unregulated right-hand priority rule for cars does not apply.
    Correct answer: A
  - **Correct Answer:** B

- **Question ID: e89edd75**
  - **Category: aerial**
  - **Image:** Figure 11
  - **Question:** In what order will the vehicles pass through the intersection shown, if the traffic lights are not working?
    A: Tram 1, Tram 2, the car;
    B: The car, Tram 1, Tram 2;
    C: Tram 1, the car, Tram 2.

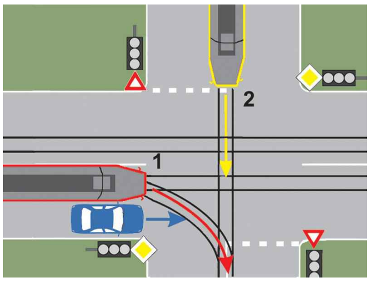


Figure 11: Image corresponding to question ID: e89edd75

  - **Reasoning:** [REASONING]**Analyzing traffic priorities**
    I'm observing the traffic signs, where yellow diamonds indicate both the east and priority direction. It's interesting that both west and east have these signs, suggesting that the road from west to east takes priority. Meanwhile, the north and south roads are minor, signifying yield signs. It appears that the main road runs west to east, with a tram and a blue car from the west. Since both want to go south, we'll need to establish the priority for who goes first. [REASONING]**Examining tram and car priority**
    I'm analyzing the situation where both a tram and a car are sharing the track. According to the rules, cars must yield to trams, especially in intersections without functioning signals. The blue car must yield to all trams. Between two trams, one is on a major road (priority) and the other on a minor road. The minor road tram must yield to the tram on the main road due to the yield sign. So, tram 2 yields to tram 1 first, then to the blue car, but the car ultimately gives way to tram 1 as well. [REASONING]**Determining tram and vehicle order**
    I'm working through how priority applies at this intersection without functioning signals. Trams have implicit priority, but with two trams, I check the signage. It seems like the signs apply to both trams. The secondary yield sign means tram 2 must yield to tram 1, which is on the main road and has no yield sign.
    So, tram 1 goes first, then tram 2, and finally the blue car, which must yield to the trams. Therefore, the correct order is tram 1, tram 2, then the car.
  - **Output:** According to Romanian road legislation, at this intersection without working traffic lights, priority signs apply first, then the rule that trams have priority over other vehicles (OUG 59/2002 art. 1) and, if in equal position, the "right-hand rule" as reflected by road signs.
    1. The priority road is the west–east axis (priority road signs on west and east); north and south are minor roads (yield signs). 2. Therefore, tram 1 (on the priority road) enters first. Also, the car on the same road cannot overtake the tram (Reg. 107 art. 3). 3. Next comes the second tram (coming from the secondary north road), after the first tram clears the intersection. Although on a minor road, it is still a rail vehicle and has priority over the car (OUG 59/2002 art. 1). 4. The car enters last, after both trams have completed their maneuvers.
    Correct answer: A
  - **Correct Answer:** B

- **Question ID: 3469375f**
  - **Category: aerial**
  - **Image:** Figure 12
  - **Question:** What is the correct way to proceed if you are driving the red car?
    A: Turning left, you may only enter the intersection last;
    B: You enter the intersection first;
    C: You must yield to the tram.
  - **Reasoning:** [REASONING]**Analyzing road intersection**
    I'm looking at an intersection with a vertical road that seems to be the main priority road, while the horizontal road might be major but not prioritized. I see two yellow diamond signs indicating that the horizontal road has priority, meaning vehicles on it have the right of way. There's a red car on the minor eastbound road with a yield sign, indicating

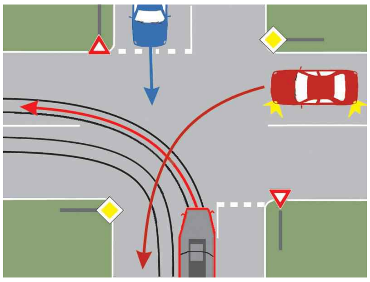


Figure 12: Image corresponding to question ID: 3469375f

it's about to turn left across the priority road, while a tram approaches from the left on tracks marked in black curves. [REASONING]**Examining tram intersection**
I'm analyzing an intersection with tram tracks represented by thick black lines. They converge from the left and diverge on the right. There are three tram tracks merging before splitting after the intersection. A red car is at the intersection, wanting to turn left across the tracks, while a blue tram travels south on the vertical road, marked with a yield sign. The horizontal road has priority signs on both sides, confirming that the red car is on the priority road, positioned on the right side of the intersection. [REASONING]**Interpreting traffic directions**
I'm observing a red car on the horizontal priority road, which intends to make a left turn. The car's direction arrows and blinking yellow signals indicate the intended turn. Although the red arrow path suggests it curves downwards to the south, it's clear the car is facing west and wants to turn left towards that direction. The car has priority on this road, and there's a yield sign nearby, but it's directed towards the bottom vertical road and not affecting the red car's right of way. [REASONING]**Evaluating intersection rules**
At this intersection, I realize the red car must yield to any tram if one is approaching and also must cross before the blue car since it's on the priority road. The correct approach for the situation seems to be yielding to the tram while having priority over the blue car. The red car has to execute a left turn, which means yielding to traffic from the opposite direction and trams, confirming the red car goes first, making B and C the right answers.

- **Output:** The driver of the red car is on the priority road (indicated by the "Priority Road" sign) and is making a left turn, while the transverse road is a non-priority road (marked with "Yield" sign).
  1. Relative to the blue car (coming from the north), you have priority, so you may enter the intersection first. – According to OUG 61 and Regulation-130, vehicles on the priority road have right of way over those on the non-priority road. 2. However, in an intersection with uncontrolled traffic, you must yield to trams traveling on rails, except those turning left (in which case the tram loses priority). – According to OUG 59 para. (1) and (2), vehicles must allow the tram to proceed before completing their maneuver.
  The correct answers are B and C.
  Correct answer: B, C
- **Correct Answer:** B

- **Question ID: a3c595fb**
  - **Category: misc**
  - **Image:** Figure 13
  - **Question:** The sign indicates that:
    A: Stopping and parking restrictions have ended;
    B: All restrictions have ended;
    C: The limited-time parking zone has ended.

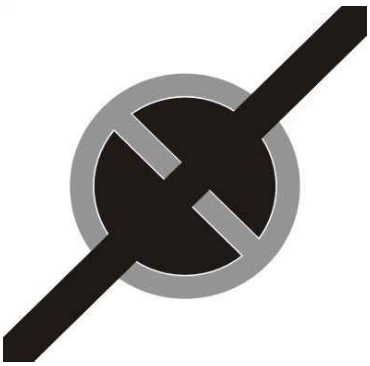


Figure 13: Image corresponding to question ID: a3c595fb

  - **Reasoning:** None
  - **Output:** According to the provisions of SIGN-95 "End of all restrictions", the sign in the image is the standard sign that announces the end of all previous road restrictions. It does not refer only to stopping and parking (option A) nor only to the end of a limited-time parking zone (option C), but to all restrictions.

Correct answer: B
- **Correct Answer:** C

- **Question ID: 37af955f**
  - **Category: misc**
  - **Image:** Figure 14
  - **Question:** The adjacent road marking means:
    A: Space designated for weighing vehicles;
    B: Pedestrian crossing marking;
    C: Marking for parallel parking.

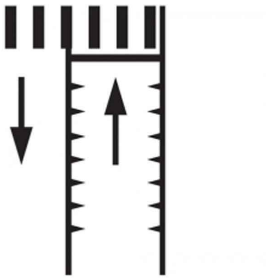


Figure 14: Image corresponding to question ID: 37af955f

  - **Reasoning:** [REASONING]**Evaluating road markings**
    I'm thinking about the markings on the road. "B" seems to refer to a pedestrian crossing, which doesn't match the description I have. "C" mentions parking bays, but again, that's not it. Now "A" suggests a space for weighing vehicles, but I'm unsure if there's a specific marking for weighbridges on public roads. There are some markings before police checkpoints, but they're not official. In the image, solid lines with triangular arrows don't suggest a reserved lane. Overall, it sounds like the correct answer might be none. [REASONING]**Analyzing road markings**
    I'm looking at the triangles along the lines. They may indicate a reserved lane for public transport, but the triangles are confusing since some point in opposite directions. In the image, it seems the central lane has arrows indicating direction, while the edges have multiple triangles directing inward. The left side shows a lane with a downward arrow, and the right has an upward arrow despite some unclear markings. The central hatched area indicates traffic separation, which doesn't match any answer choices. Option A mentions a weight station, but I'm unsure if that applies here. [REASONING]**Determining the correct answer**
    The exam requires a letter response, but it seems none of the options fit. The markings indicate a prohibited lane, not meant for parking, pedestrian crossings, or weighing. So, it seems like the correct choice should be "none." However, the options are limited to A, B, or C. If I must choose, I'd say "none," and I can format it as "Raspuns corect: niciuna." But since the instructions say to write only letters, I'll note that carefully.
  - **Output:** According to the Road Signaling Regulation, the marking consisting of two parallel longitudinal lines enclosing an area filled with oblique lines ("triangles") does not indicate parking, pedestrian crossing, nor a weighing device. This type of marking falls under "various – spaces prohibited to traffic" (Reg.-76, lit. d)2 and Reg.-83 para. (4)), delimiting an area into which vehicles are forbidden to enter.
    None of options A, B, or C matches the meaning of the marking in the image.
    Correct answer: none
  - **Correct Answer:** B

- **Question ID: 47e15cb4**
  - **Category: misc**
  - **Image:** Figure 15
  - **Question:** What information does the sign in the image provide?
    A: An uneven passage follows, 3 km from the tunnel;
    B: A tunnel follows, 3 km from the sign;
    C: The upcoming tunnel is 3 km long.
  - **Reasoning:** [REASONING]**Analyzing road signs**
    I'm looking at the official drawings for signs. The sign C.2b seems to indicate something about distance, specifically regarding a tunnel 3 km away. I notice that A refers to an uneven passage which doesn't fit because the icon shows a tunnel, not humps. C mentions the length of the tunnel as 3 km, but if that were a dangerous section, additional information would be required. This suggests the main sign indicates general information rather than a warning. [REASONING]**Justifying the sign interpretation**
    I'm analyzing the signs and concluding that it's not about the length of a dangerous section, which means C is incorrect. The right answer is B. The definitions indicate that the distance sign is used under a warning sign, but here we have an informative sign. The type C.2 sign shows the distance to a tunnel, so it accurately indicates that the tunnel is 3 km ahead. I've evaluated options A and C, and they don't fit the criteria, reinforcing that B is the correct answer.
  - **Output:** According to the Road Signaling Regulation:
    1. The image shows an information sign ("category C – orientation and information signs") placed on a blue background, with a tunnel symbol over a white background and below, the inscription "3 km". 2. This type does not indicate the length of a dangerous section (this would be shown by an additional "Dangerous section length" panel (SIGN-280)

Figure 15: Image corresponding to question ID: 47e15cb4

mounted under a triangular warning sign), but rather the distance to the indicated object. 3. Therefore, the meaning of the sign is: "a tunnel follows, 3 km from the location of this sign."
Options A and C are incorrect: - A: The symbol does not indicate an uneven passage (that would be a different icon), but a tunnel. - C: The length of the tunnel is not 3 km – there is no danger length panel, only an indication of distance to the tunnel.
Correct answer: B
  - **Correct Answer:** C

- **Question ID: af30ec04**
  - **Category: pov**
  - **Image:** Figure 16
  - **Question:** Have you selected the correct lane to turn left at the upcoming intersection?
    A: Yes;
    B: No;
    C: You are required to go straight only.

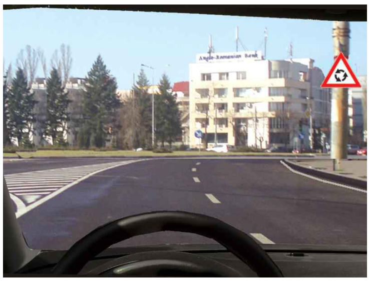


Figure 16: Image corresponding to question ID: af30ec04

  - **Reasoning:** [REASONING]**Identifying road signs**
    I'm noticing an intersection ahead with what looks like a roundabout sign. To the right, there's a triangular sign with black arrows in a circle, which serves as a pre-warning for the roundabout. It's defined as SIGN-40, indicating a roundabout is ahead. After that, we should see the actual mandatory blue circular sign with white arrows at the intersection. I think the triangular sign indicates the upcoming roundabout, while the blue sign directs us on how to navigate it. [REASONING]**Interpreting the question about turning left**
    The question asks if I've chosen the correct lane to turn left at the upcoming intersection. It could imply I'm approaching a roundabout and need to consider how roundabouts work. The triangular sign is a warning for a roundabout ahead. If I see a blue sign with white arrows, it indicates it's indeed a roundabout. In such cases, you typically circulate on the right, so turning left isn't standard. However, there could be confusion about whether it's a normal intersection or a circular one. [REASONING]**Evaluating lane selection**
    If I'm exiting leftish (the second exit) at a roundabout, I need to ensure I'm in the proper lane. In Romania, I should pick my lane based on the exit direction. The leftmost lane is reserved for left turns, while the central lane is for going straight. Since I'm currently in the middle lane, it's incorrect for a left turn. There's no sign mandating I go straight, so my answer is that I can't make that turn from my current lane. Thus, the answer must be B.
  - **Output:** According to art. 14 para. (1) and (2) of the Regulation for implementing OUG 195/2002 and usual practice: 1. In the absence of arrow markings or direction signs designating turn directions for each lane, to turn left, drivers must move to the lane adjacent to the road axis (the leftmost lane), at least 50 m before the intersection. 2. In the image, the vehicle is in the second lane (of three), so not in the required position for a left turn (not in the leftmost lane). 3. There is no "Straight ahead only" sign (SIGN-109) that would prohibit turning left and require traveling only straight ahead.
    Therefore, the chosen lane is not correct for a left turn.
    Correct answer: B
  - **Correct Answer:** A

- **Question ID: a64a8f4c**

- **Category: pov**
- **Image:** Figure 17
- **Question:** What is the correct way to proceed in the presented situation?
  A: You are required to wait;
  B: You have no obligation;
  C: You signal to the yellow bus that it may proceed.

Figure 17: Image corresponding to question ID: a64a8f4c

- **Reasoning:** [REASONING]**Analyzing traffic obligations**
  I'm considering a scenario where a bus is stationary and preparing to leave the bus stop while I'm approaching a zebra crossing. I need to figure out whether I have to wait for the bus, if I have no obligation, or if I should signal for it to proceed. According to legislation, drivers next to the edge must give priority to buses trying to re-enter traffic and ensure they don't endanger others. So, it seems I must wait. [REASONING]**Determining traffic priorities**
  I'm analyzing the rules regarding a bus at a stop preparing to re-enter traffic. The legislation states that I must yield to public transport vehicles that signal their intent to return to the road. Essentially, I need to reduce speed or stop to let the bus in. The options suggest I either wait or signal the bus, but I don't actually need to signal; yielding is enough. So, the correct choice is definitely A: I'm obliged to wait.
- **Output:** The driver of a vehicle traveling in the lane next to the edge of the carriageway is required to give priority to public transport vehicles when they signal their intention to re-enter traffic from a bus stop with an indentation and have ensured that their maneuver does not endanger the safety of other road users (OUG 43/1997 art. 2 para. (2) and Regulation 131/2010 art. 1 para. (1)). Therefore, when the bus in the image signals its exit from the station, you must reduce speed and, if necessary, wait to give it priority.
  Correct answer: A
- **Correct Answer:** B

• **Question ID: 1b4d81f6**

- **Category: pov**
- **Image:** Figure 18
- **Question:** The sign prohibits access to:
  A: Cars, motorcycles without sidecars, and animal-drawn vehicles;
  B: Motor vehicles and animal-drawn vehicles;
  C: Cars, motorcycles, and hand-pushed or hand-drawn vehicles.

Figure 18: Image corresponding to question ID: 1b4d81f6

- **Reasoning:** [REASONING]**Interpreting sign imagery**
  I'm seeing a motorcycle silhouette, which has two wheels and a rider shape. There's also what I think is a bicycle icon in the bottom left, showing pedals and handlebars. The bottom right could be a horse-drawn carriage. It seems the sign prohibits cars and motorcycles, as well as animal-drawn vehicles. The specific sign design looks like it divides into sections with three icons: a car, possibly a motorcycle, and something else, but definitely no bicycle. [REASONING]**Identifying the sign type**
  It seems likely that this is SIGN-64, which prohibits access to vehicles except for motorcycles without sidecars. Typically, that sign has a car silhouette and a motorcycle silhouette, often stacked. From what I see, there appear to be three images: at the top is a car, followed by a motorcycle or moped, and then a horse-drawn carriage at the bottom. The bottom left looks like a bicycle with pedals, while the right shows a horse with a cart. So the sign probably covers cars and horse-drawn vehicles but not bicycles. [REASONING]**Analyzing sign categories**

The usual signs like SIGN-64, SIGN-75, and others prohibit various types of vehicles, but I notice this sign may group three categories, which is uncommon. The bottom right image seems to show a handcart, not a horse-drawn vehicle. So, it depicts a car, motorcycle, and hand-pushed cart. Given what I see, this suggests it prohibits cars, motorcycles, and hand-pushed vehicles, matching option C. The absence of animal imagery strengthens this conclusion, so I'll choose answer C.

– **Output:** According to the Regulation for implementing OUG 195/2002, the prohibition sign includes in its pictograms a car, a motorcycle, and a cart/hand-cart (vehicle pushed or drawn by hand), so it prohibits access precisely for these three categories of vehicles.
  – It does not refer to animal-drawn vehicles (they would be represented by a cart drawn by a horse/oxen). – There is no exception for motorcycles without sidecars (which would be signaled differently).
  Correct answer: C
– **Correct Answer:** B

F Detailed Error Analysis

F.1 Information Retrieval

Looking more in depth, if we compare strategies (1) and (3) in detail per category (see Figures 19 and 20), we can see the strengths of the first strategy. In some categories, while recall went higher on the training split, as these experiments don't involve training, the score went down on the test split. This could also be due to their small size, and having it split across so many categories decreases the number of samples per category even more.

On (6) from Figure 21, where we employ fine-tuning, we can see an overall improvement in all categories. We observe that categories such as *Highway Driving* and *Defensive Driving* underperform on the test set, due to a smaller share of the dataset, which would make it hard to learn relevant features. Similarly, the test set score is also smaller in *Sanctions and Offenses* and *General Rules*. A fully comparative analysis between all the strategies can be seen in Figures 23 and 24.

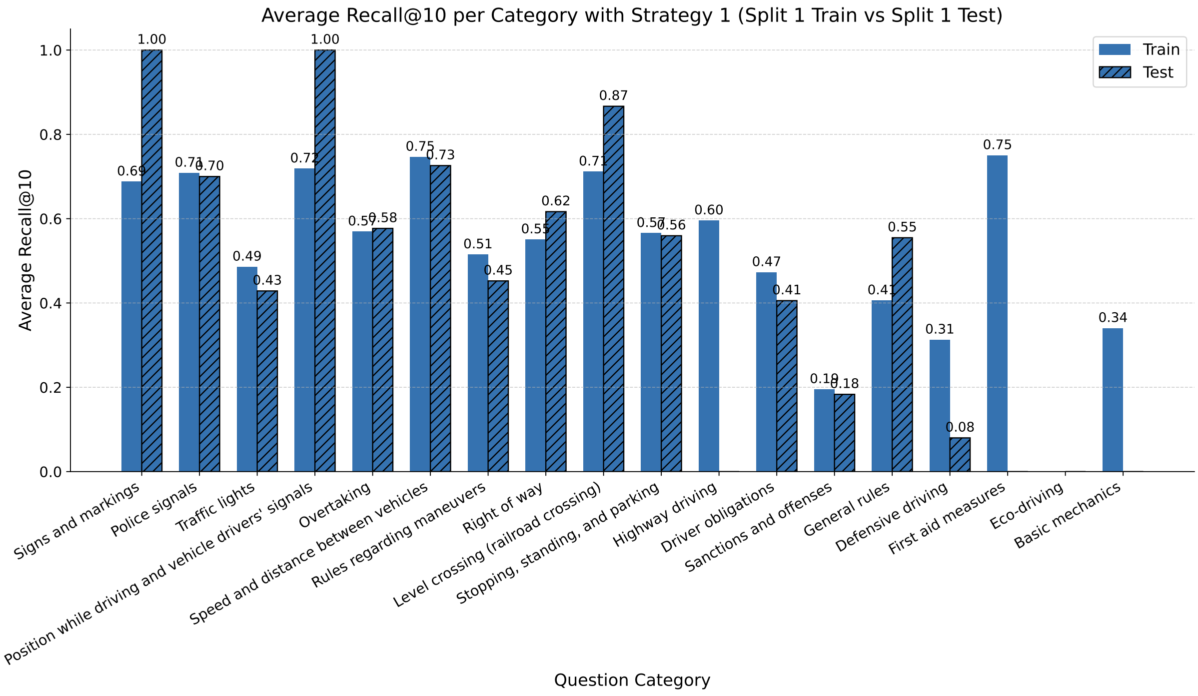


Figure 19: Recall@10 Analysis for IR task with strategy 1

F.2 Question Answering

Looking at the performances per category, in Figures 25, 26 and 27, we can see that in most categories the RAG ablation has a significant impact or in some cases similar performances. However, in "Eco Driving" and "Basic Mechanics", categories which don't need legal groundings but rather general other kinds of knowledge, the RAG is downgrading the performance (due to context bloat). But there a not a lot of entries in these categories in the Split 1 train and test. In Split 2, where we have more entries, we

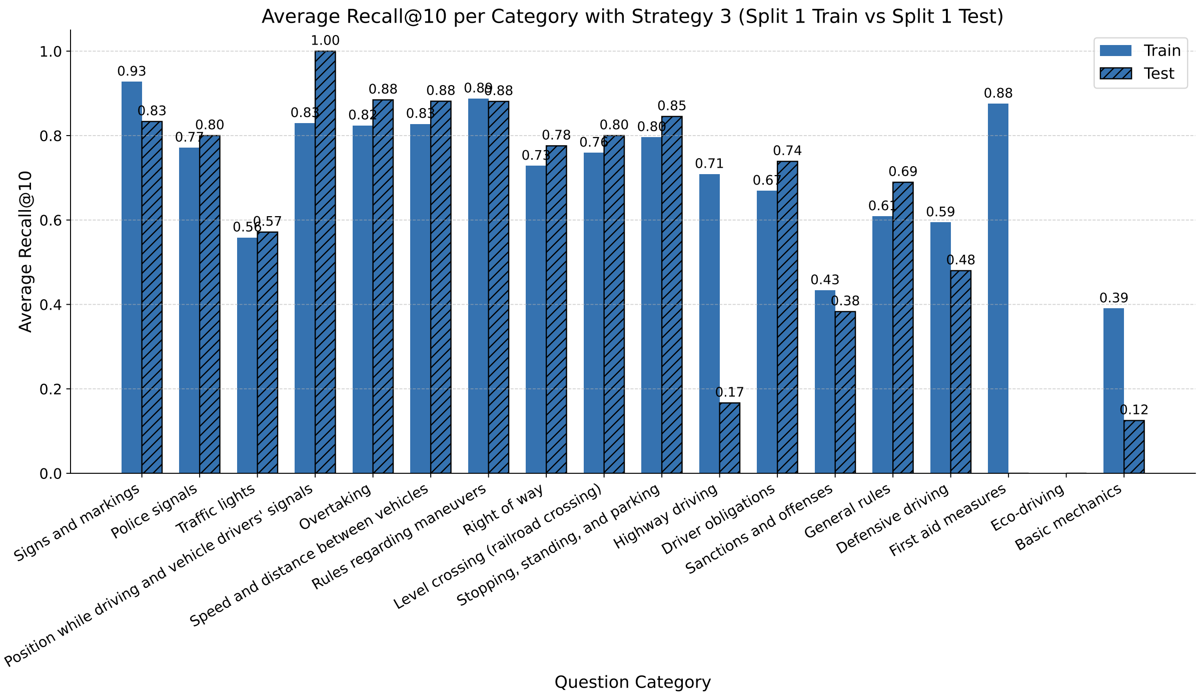


Figure 20: Recall@10 Analysis for IR task with strategy 3

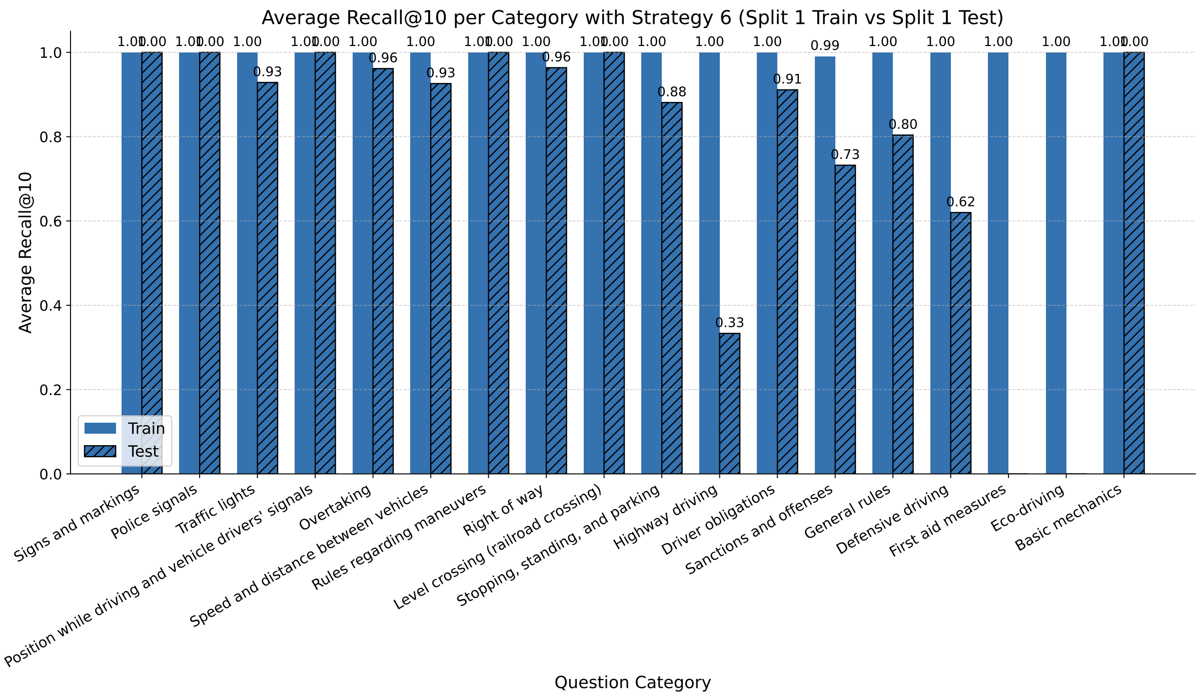


Figure 21: Recall@10 Analysis for IR task with strategy 6

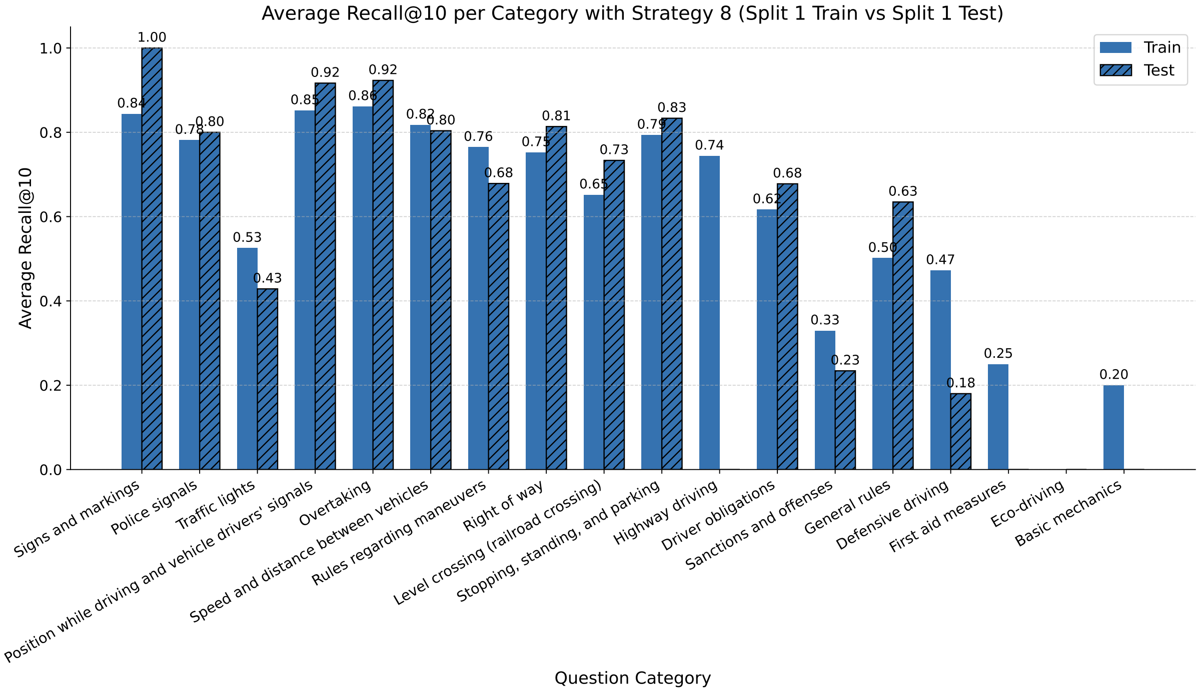


Figure 22: Recall@10 Analysis for IR task with strategy 8

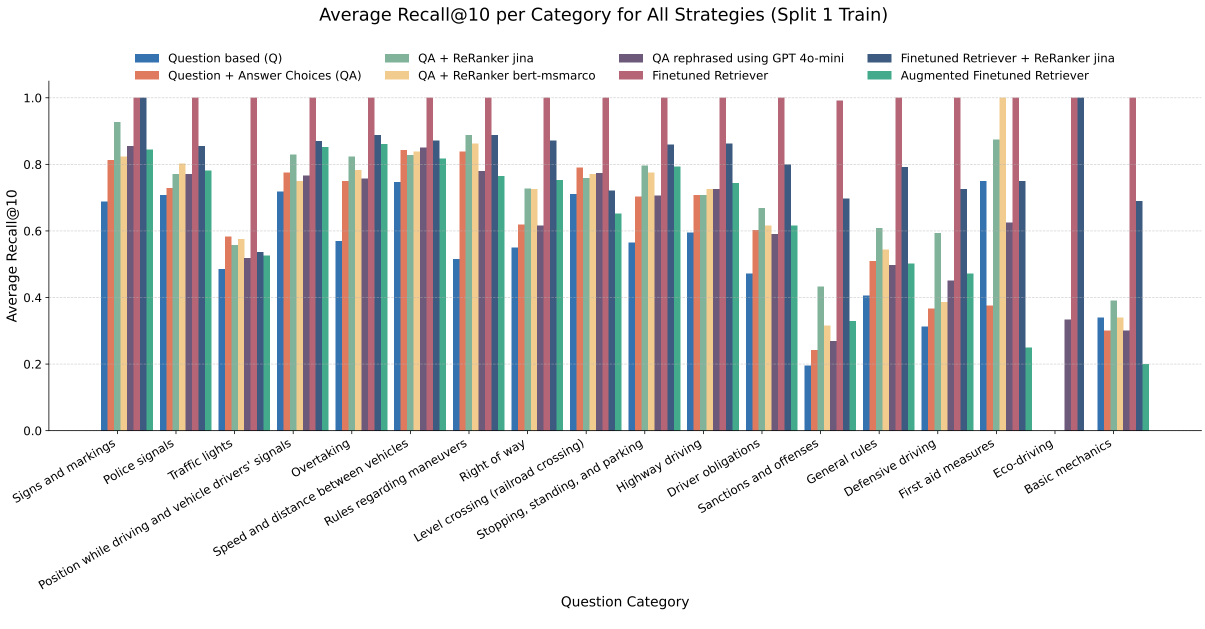


Figure 23: Recall@10 Analysis for IR task with all strategies on Split 1 Train

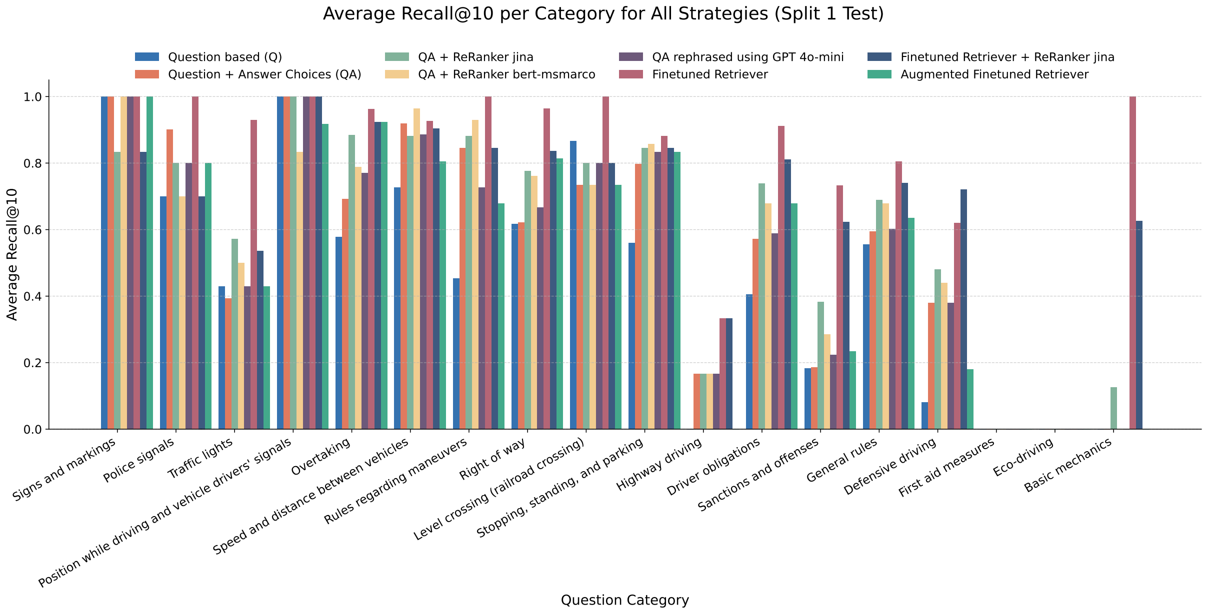


Figure 24: Recall@10 Analysis for IR task with all strategies on Split 1 Test

have similar or closer results between the strategies. We can also see that the Ideal RAG compared to our proposed solution isn't making a significant impact, showing that better retrieval doesn't lead to better QA in our case. We can see that the reasoning models perform better than normal ones, and the RAG is still making an impact in most of the categories.

For the Mistral model, we can see in Figures 28, 29, and 30 that not employing RAG helps in most of the categories. This is likely a flaw of the Mistral model, as its outputs also tend to respond in the wrong format or fail to respect instructions, instead continuing to write article paragraphs rather than responding to the query.

If we look at the number of selected answers, in Figures 34, 35 and 36, most of the time the models selects either the right amount of answers or over the amount of answers, but in very few cases less than the amount. Similarly to the recall vs precision comparison, these experiments validate that the model mostly chooses the right answers but sometimes picks more than it should when it makes errors.

If we look at the number of reasoning steps, in Figures 37, 38 and 39, when we don't include RAG, the model needs more steps to arrive at a final answer. This seems like a normal behavior, because in the RAG case it gets missing information instead of reasoning about it. We can see it needs most of the reasoning steps in eco-driving or first-aid, where law documents don't really help.

F.3 Visual Information Retrieval

Per category analysis of retrieved laws, in Figure 42, we can see it performs worse on *signs and markings*, *general rules*, and *defense driving*. The first one is more traffic sign-intensive, so it implies that the model focuses more on the signs instead of the laws related to the question.

F.4 Visual Question Answering

Looking at Figures 43 and 44, we see the model has lower performance in *Position while driving as vehicle drivers' signals* and *Stopping, standing and parking*. If we look at secondary categories, in Figures 45 and 46, we can see worse performance in the *Aerial* and similar performance in the other two. Similarly, this is the category where, in Figures 51 and 52, the models tend to use more reasoning steps than the other. If we look at the number of selected answers, in Figures 47 and 48, the model tends not to select more than enough answers (even if incorrect). We can see the comparison of reasoning steps over the primary category in Figures 49 and 50.

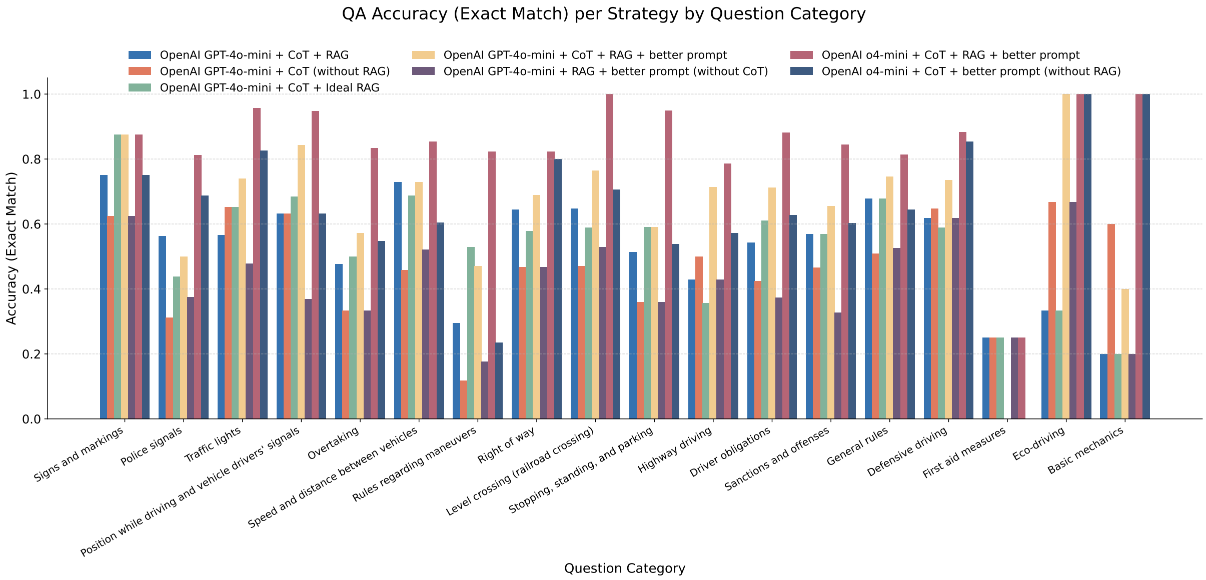


Figure 25: OpenAI models Exact Match score for QA per Strategy and Category on Split 1 Train

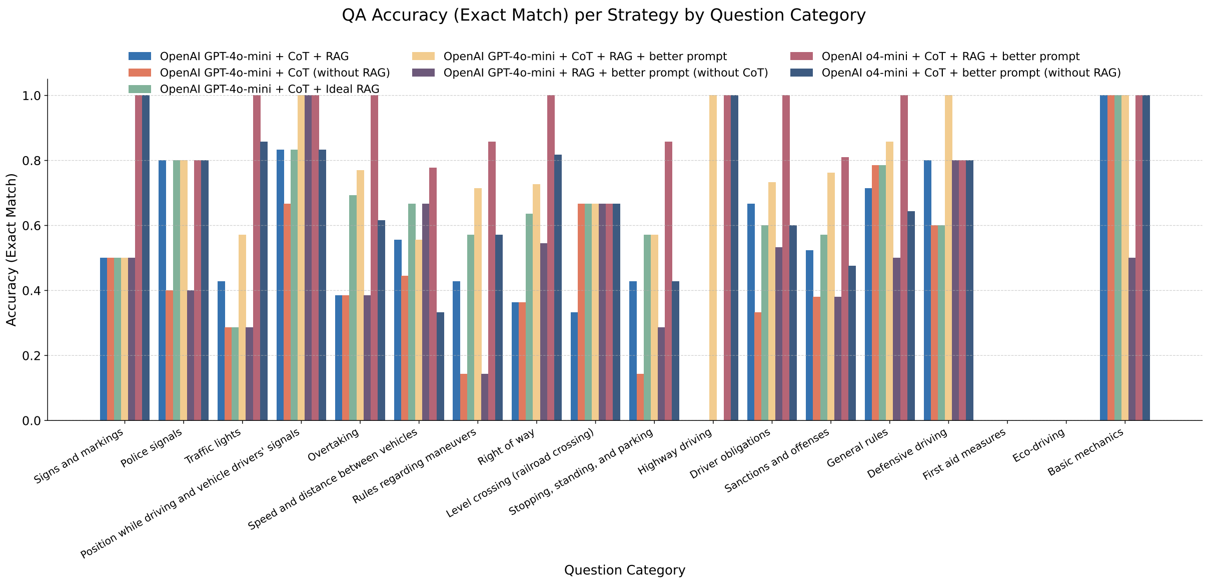


Figure 26: OpenAI models Exact Match score for QA per Strategy and Category on Split 1 Test

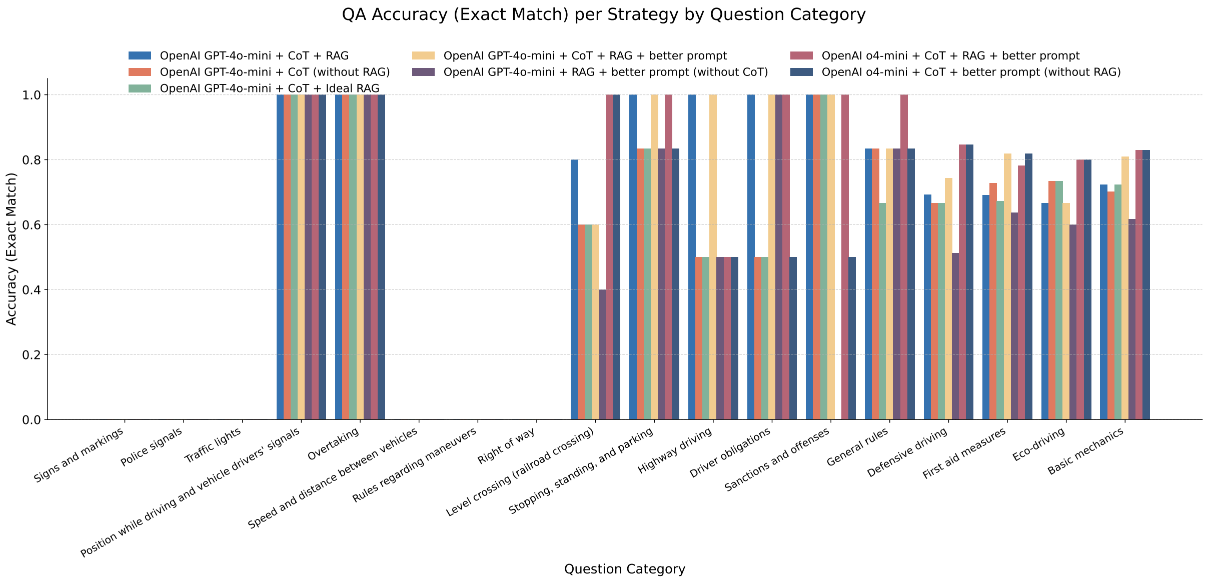


Figure 27: OpenAI models Exact Match score for QA per Strategy and Category on Split 2

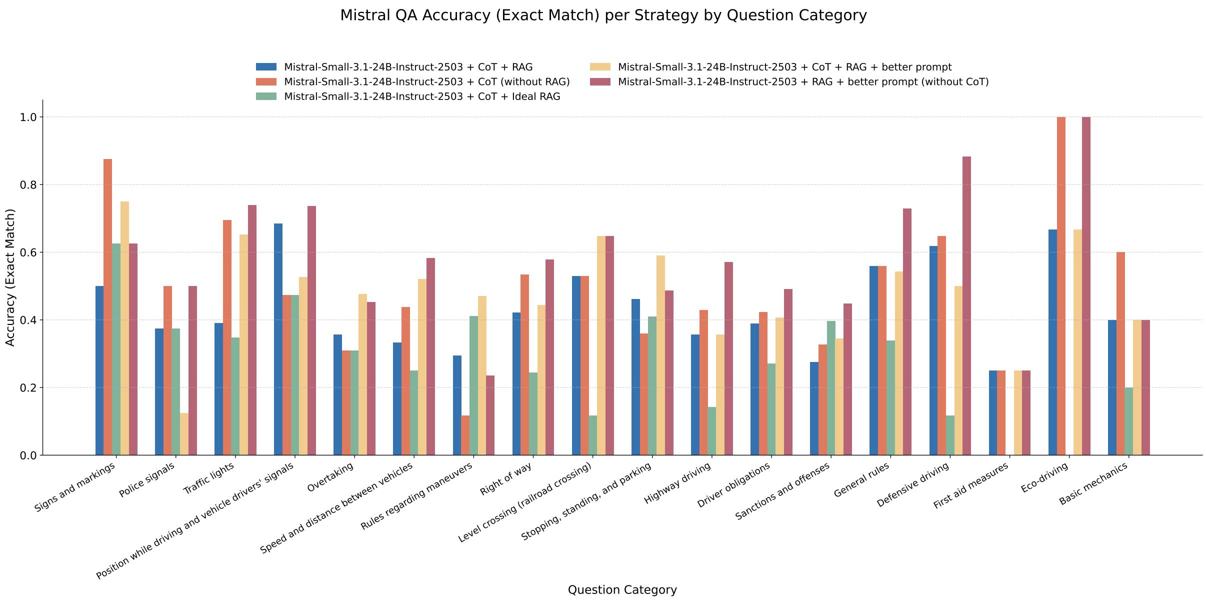


Figure 28: Mistral model Exact Match score for QA per Strategy and Category on Split 1 Train

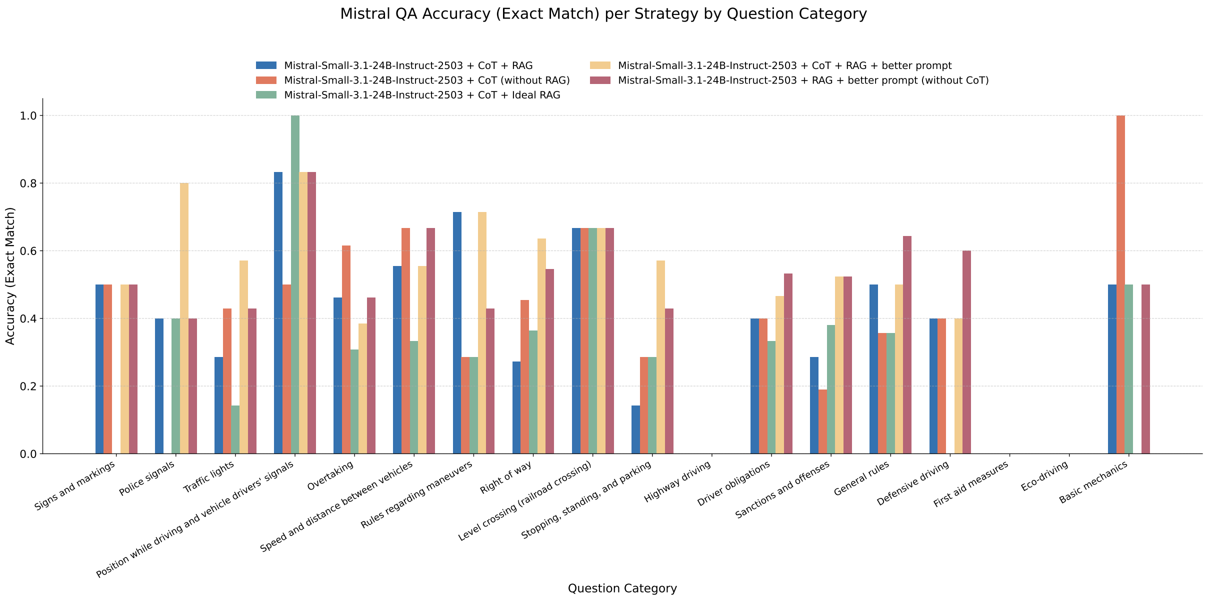


Figure 29: Mistral model Exact Match score for QA per Strategy and Category on Split 1 Test

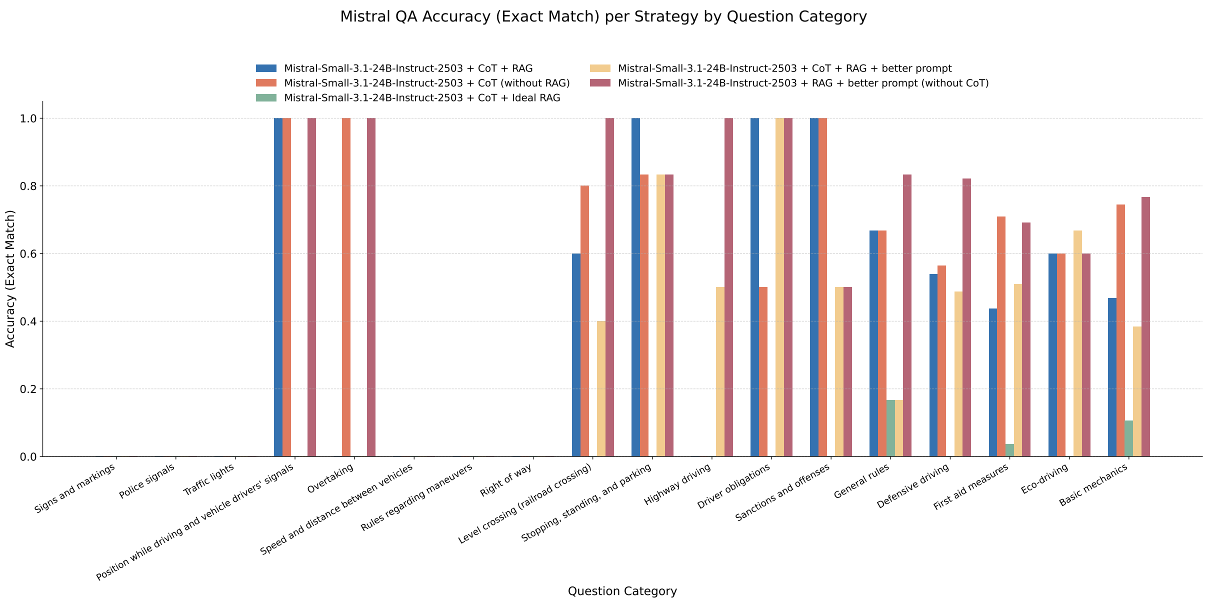


Figure 30: Mistral model Exact Match score for QA per Strategy and Category on Split 2

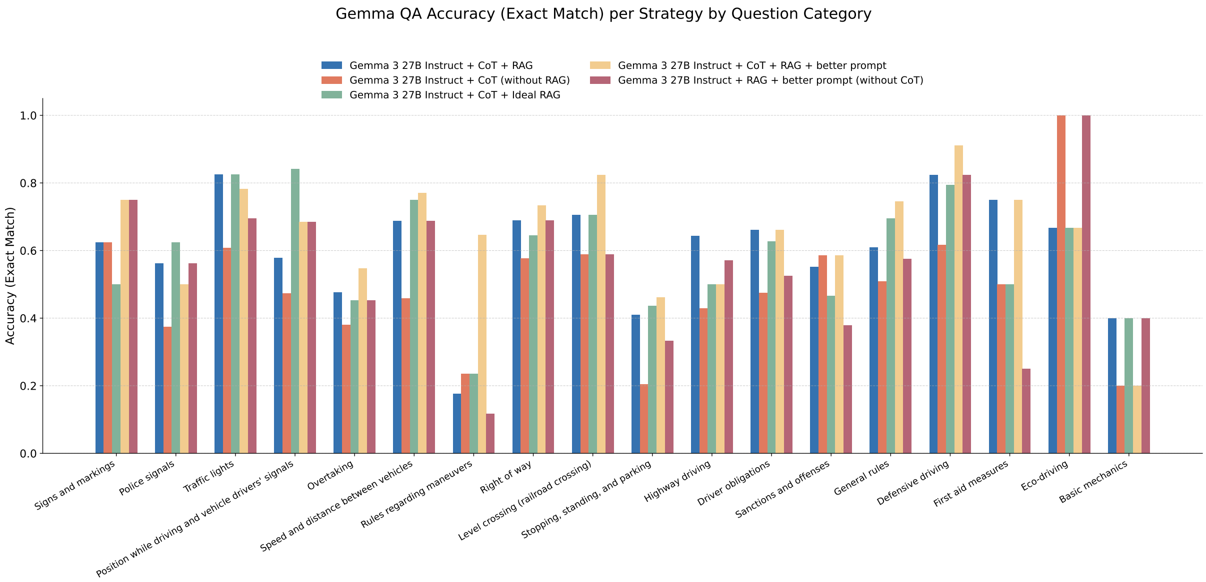


Figure 31: Gemma model Exact Match score for QA per Strategy and Category on Split 1 Train

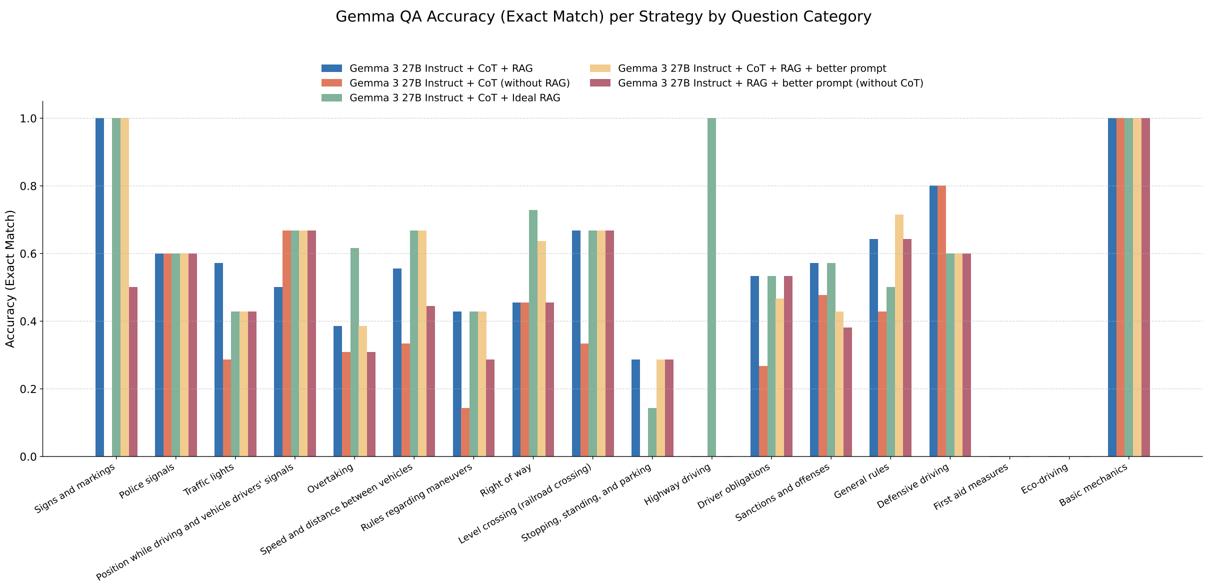


Figure 32: Gemma model Exact Match score for QA per Strategy and Category on Split 1 Test

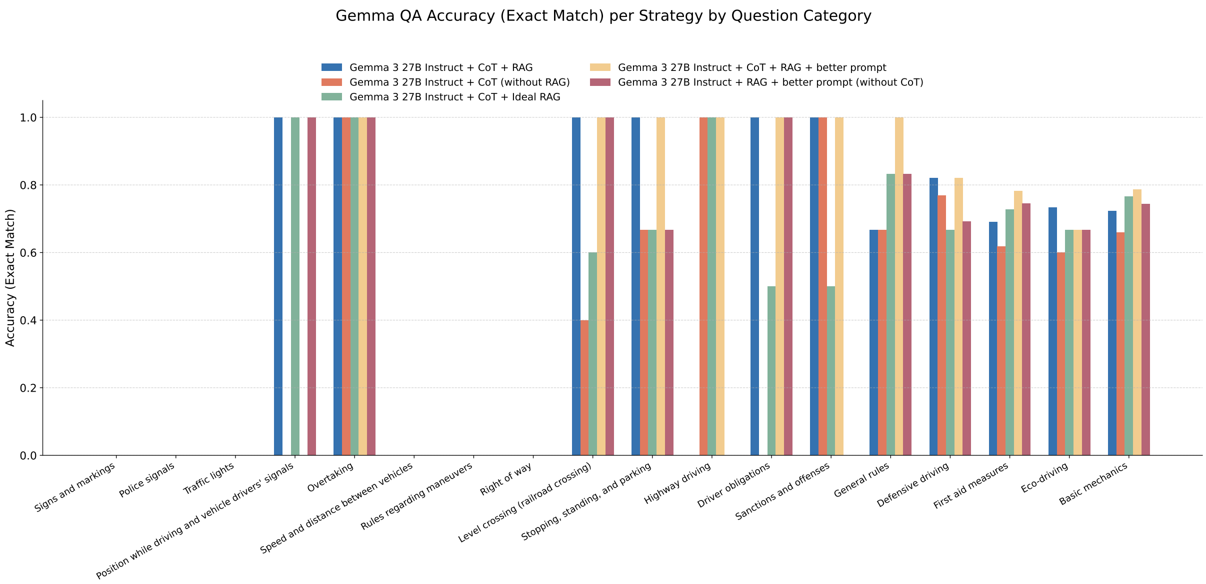


Figure 33: Gemma model Exact Match score for QA per Strategy and Category on Split 2

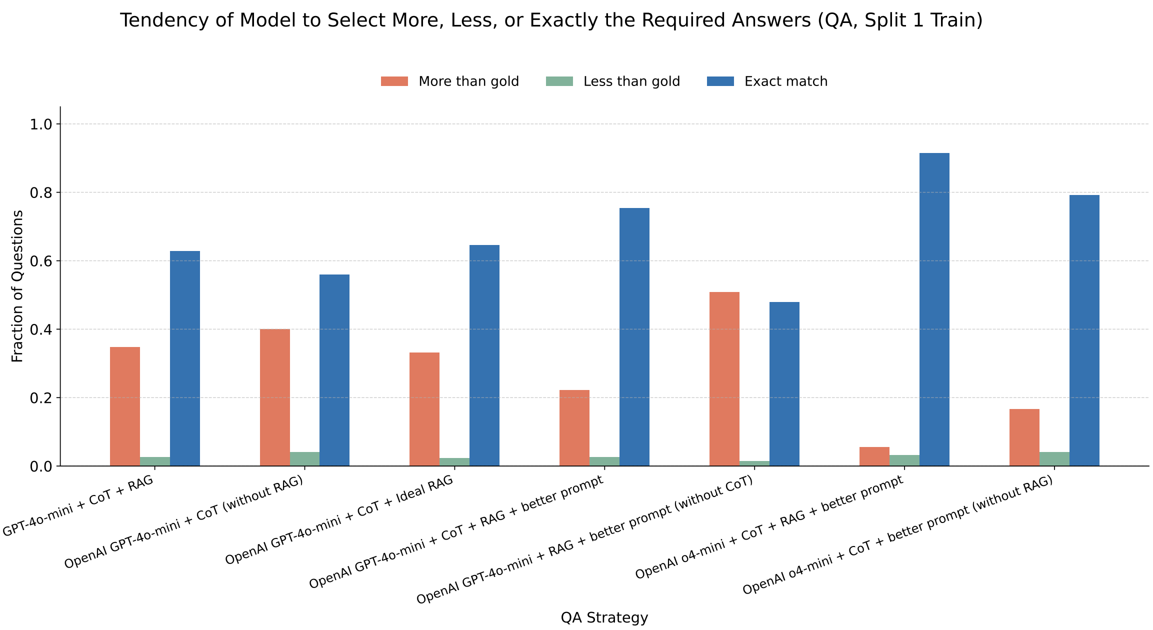


Figure 34: Tendency of OpenAI Models to Select More, Less, or Exactly the Required Answers (QA, Split 1 Train)

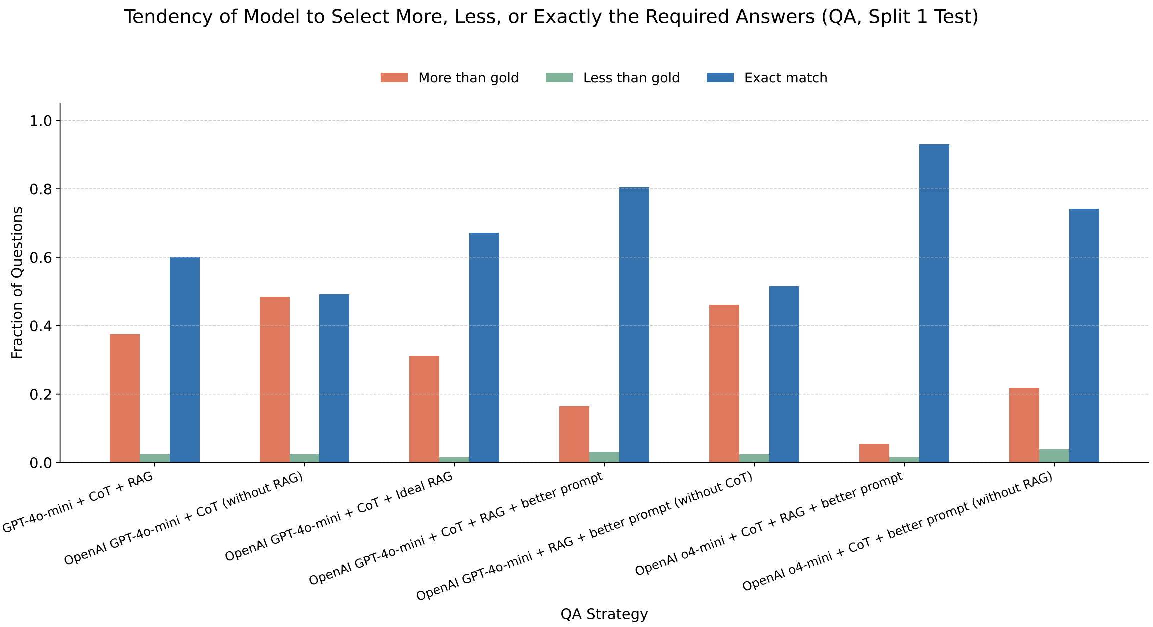


Figure 35: Tendency of OpenAI Models to Select More, Less, or Exactly the Required Answers (QA, Split 1 Test)

Figure 36: Tendency of OpenAI Models to Select More, Less, or Exactly the Required Answers (QA, Split 2)

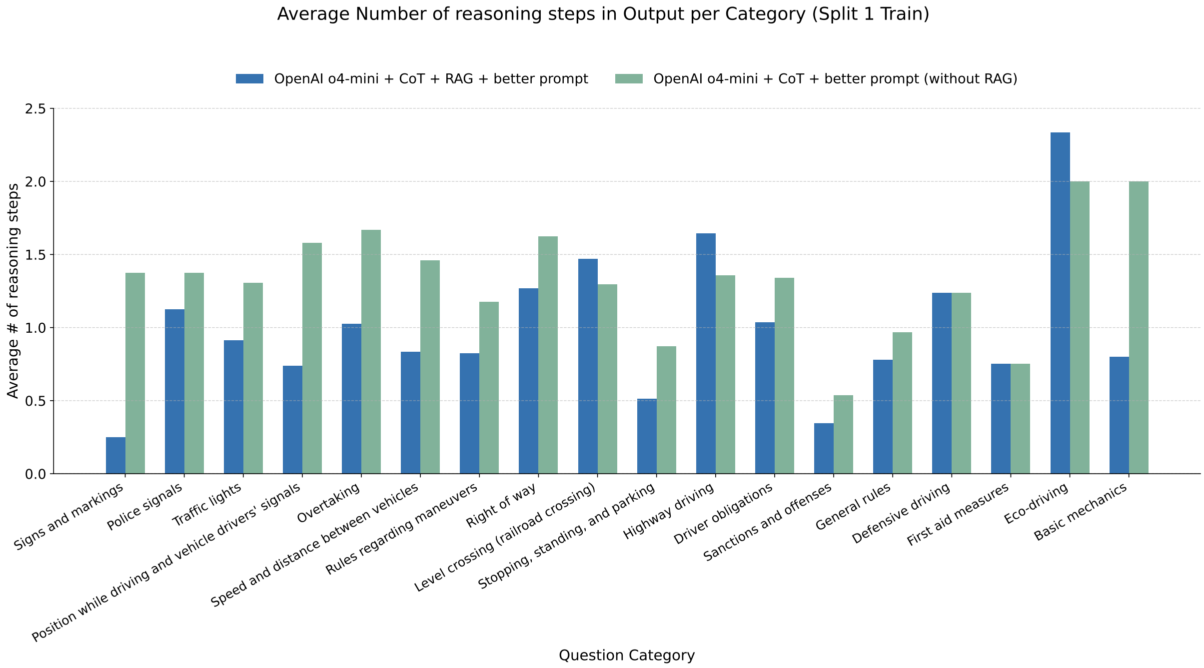


Figure 37: Average Number of reasoning steps in Output per Category (QA, Split 1 Train)

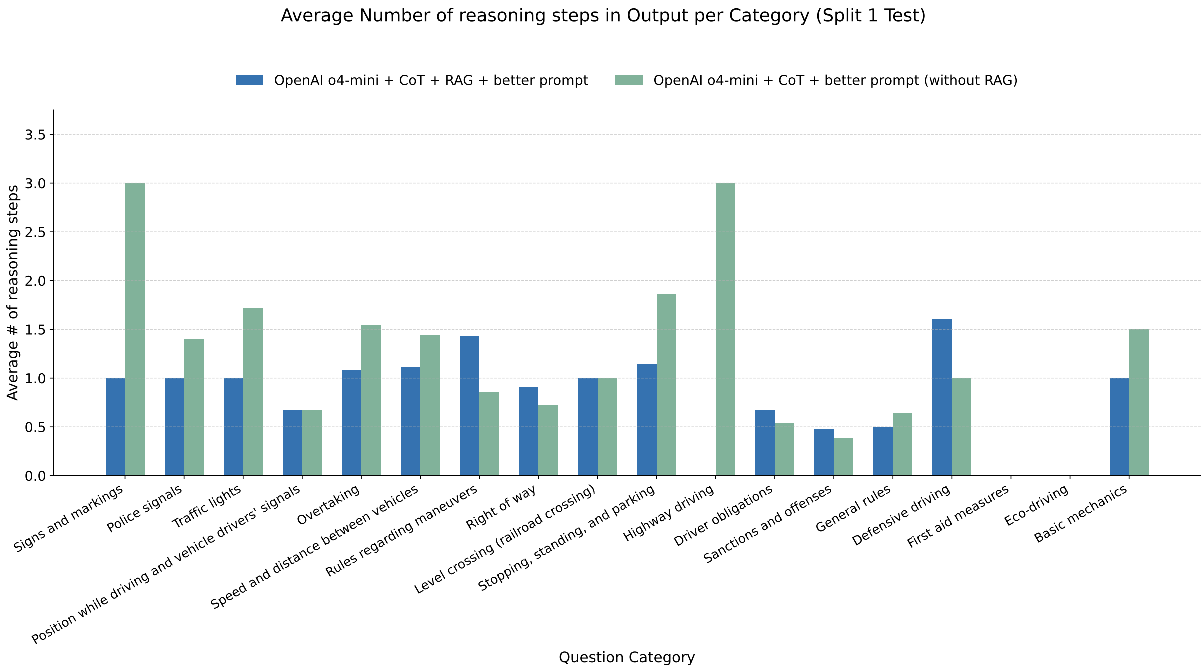


Figure 38: Average Number of reasoning steps in Output per Category (QA, Split 1 Test)

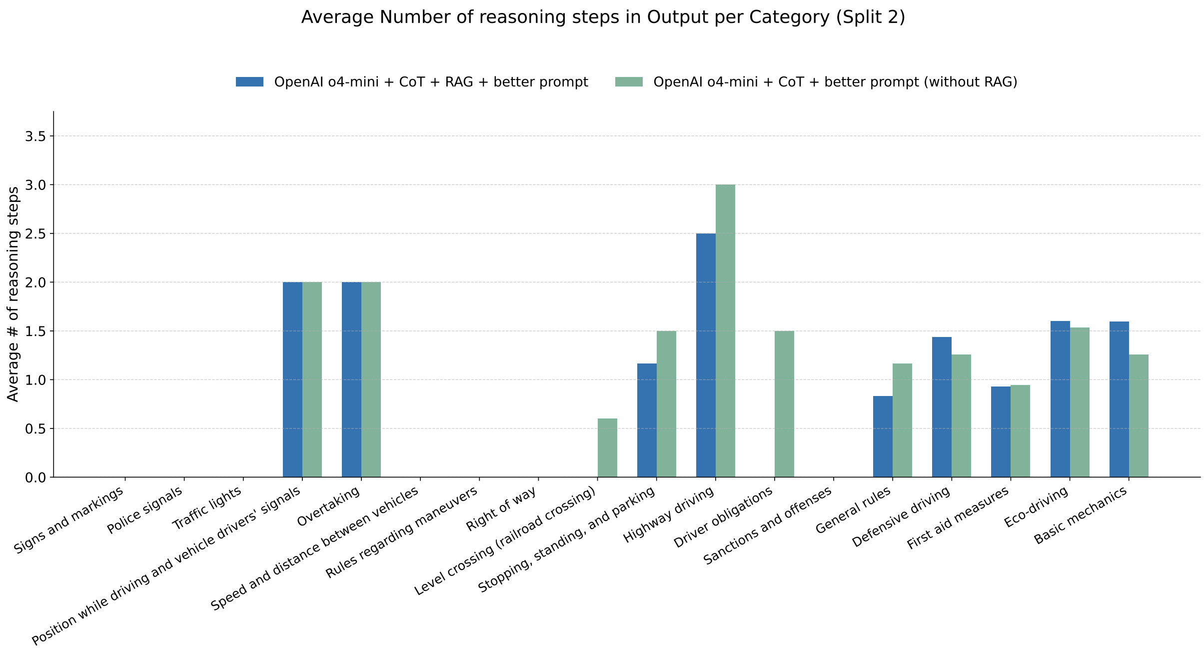


Figure 39: Average Number of reasoning steps in Output per Category (QA, Split 2)

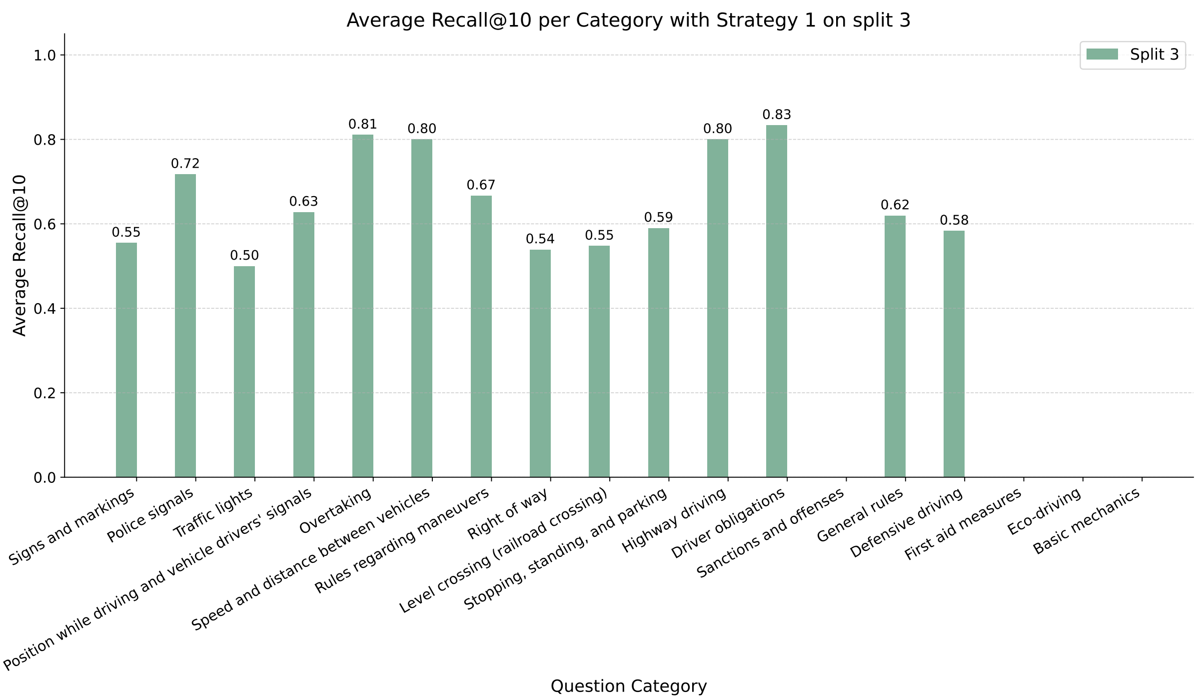


Figure 40: Recall@10 Analysis for VIR task with strategy 1 (laws)

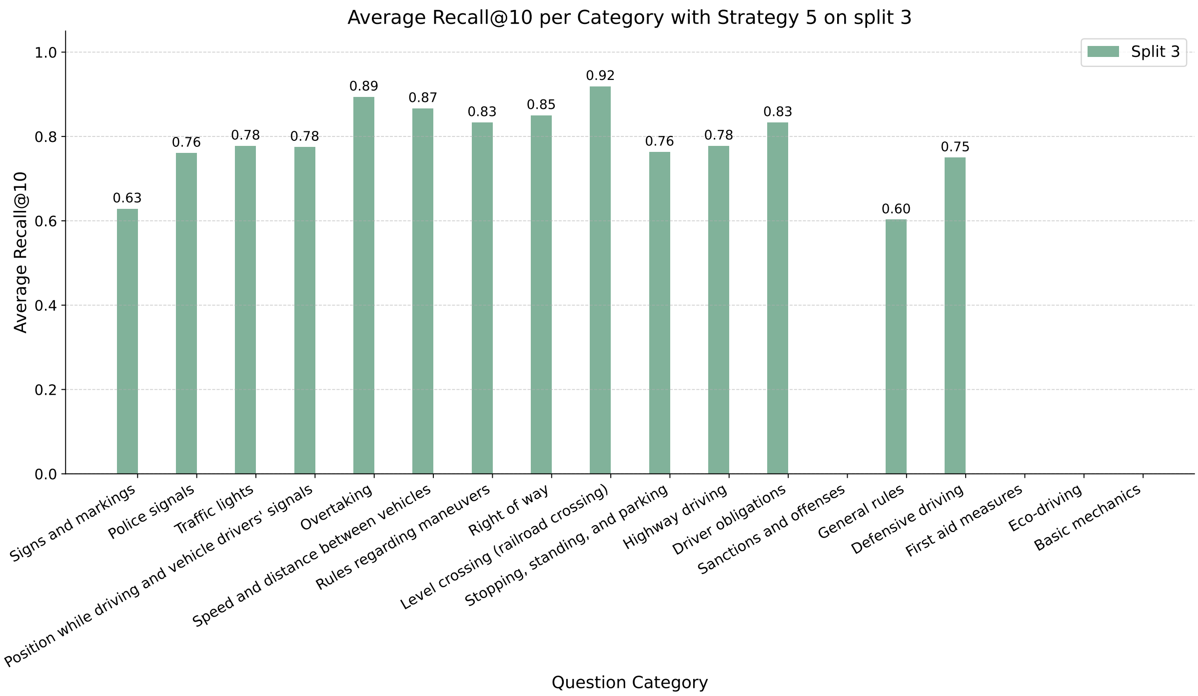


Figure 41: Recall@10 Analysis for VIR task with strategy 5 (laws)

Figure 42: Recall@10 Analysis for VIR task with all strategies on Split 3 (laws)

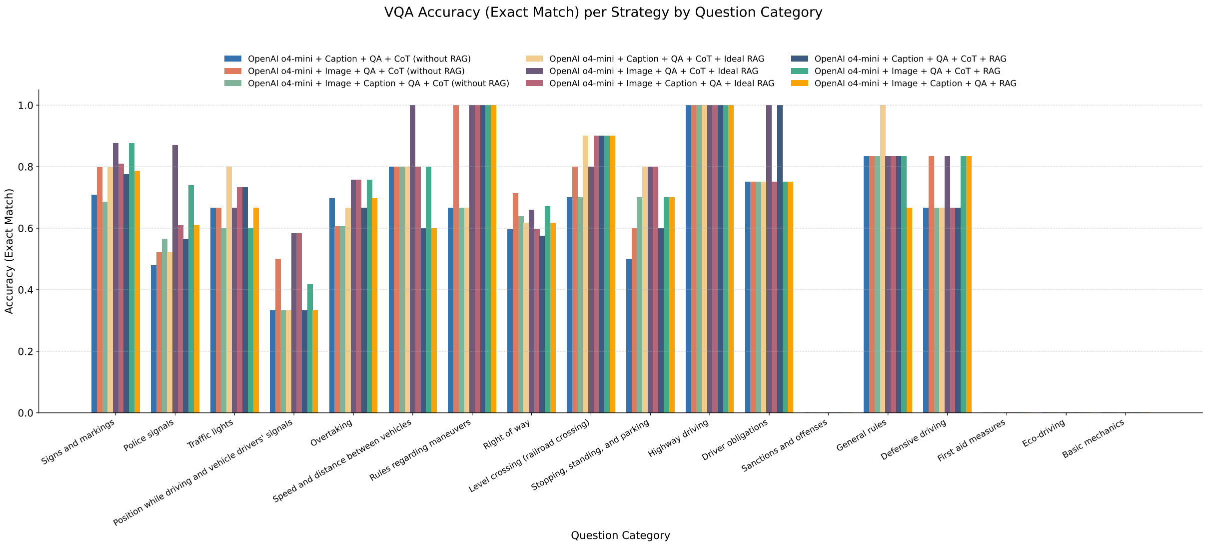


Figure 43: Exact Match score for VQA per Strategy and Category on Split 3

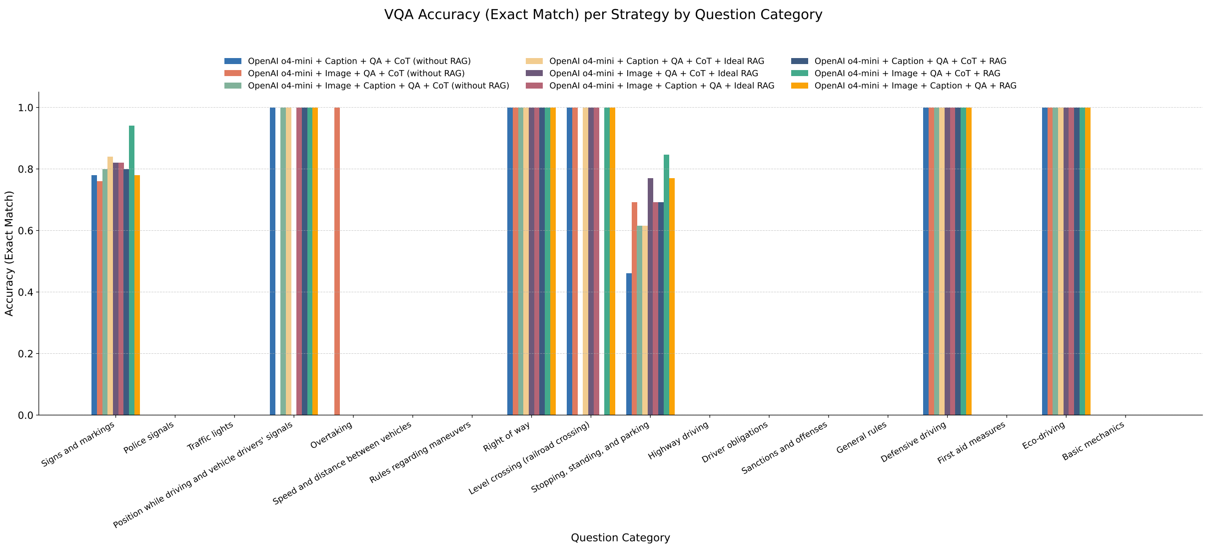


Figure 44: Exact Match score for VQA per Strategy and Category on Split 4

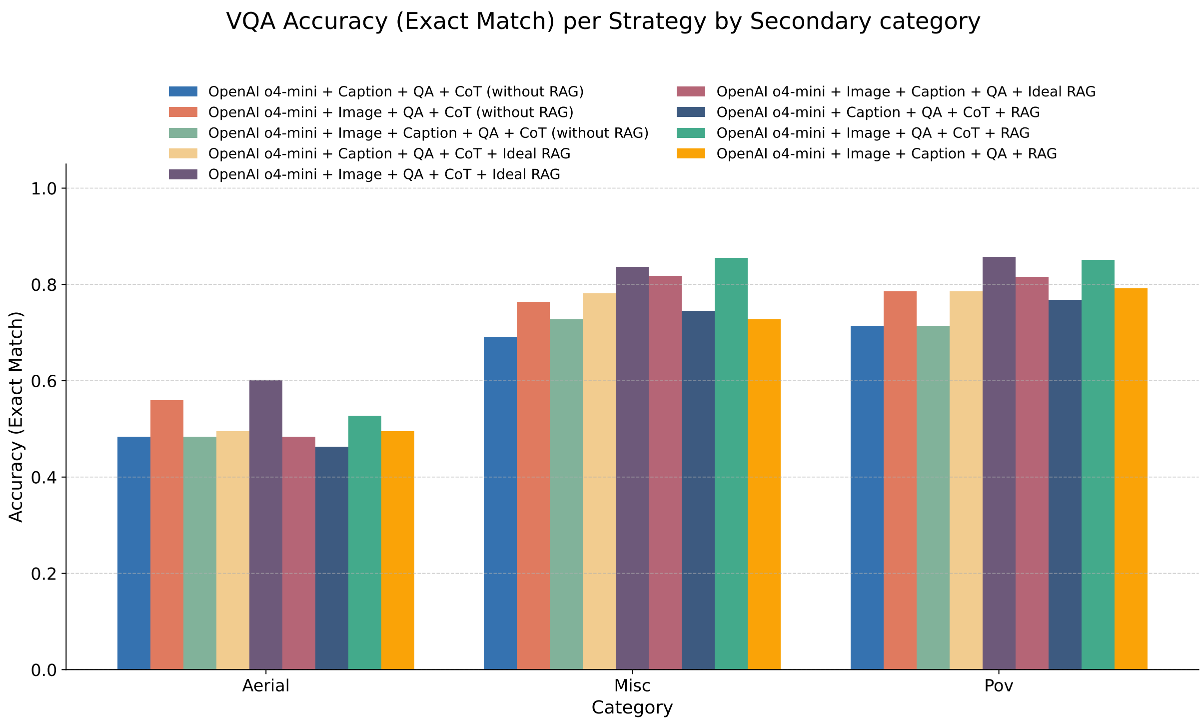


Figure 45: Exact Match score for VQA per Strategy and Secondary Category on Split 3

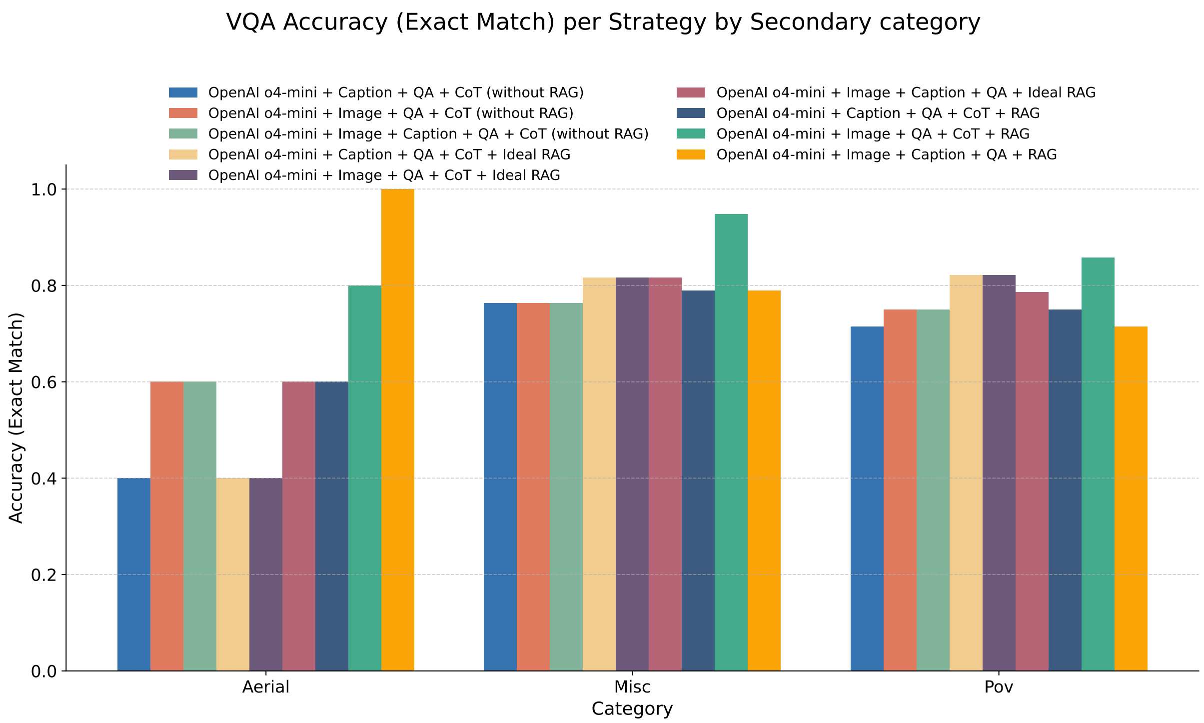


Figure 46: Exact Match score for VQA per Strategy and Secondary Category on Split 4

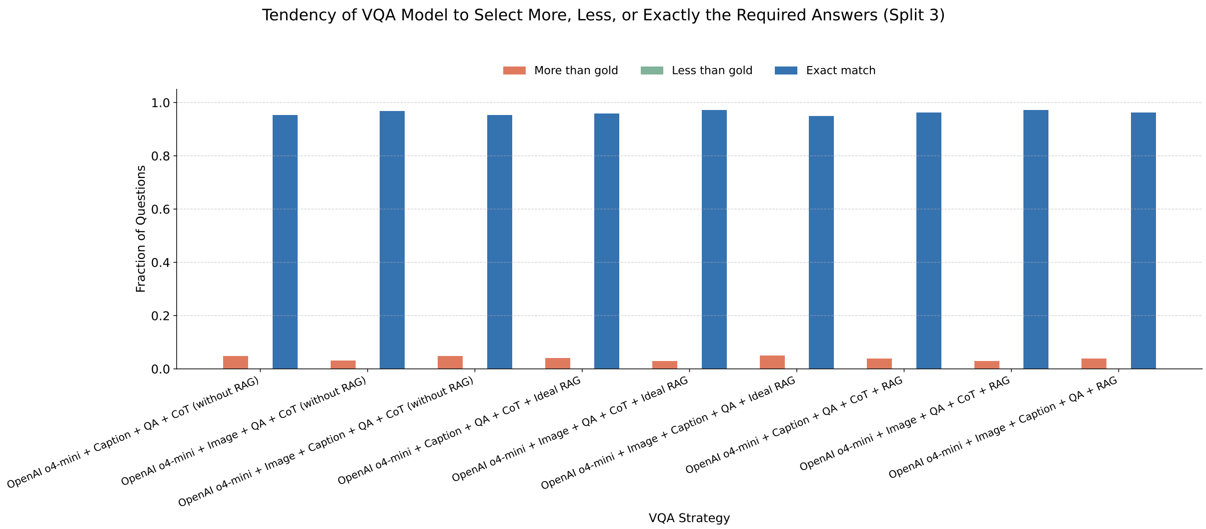


Figure 47: Tendency of OpenAI Models to Select More, Less, or Exactly the Required Answers (VQA, Split 3)

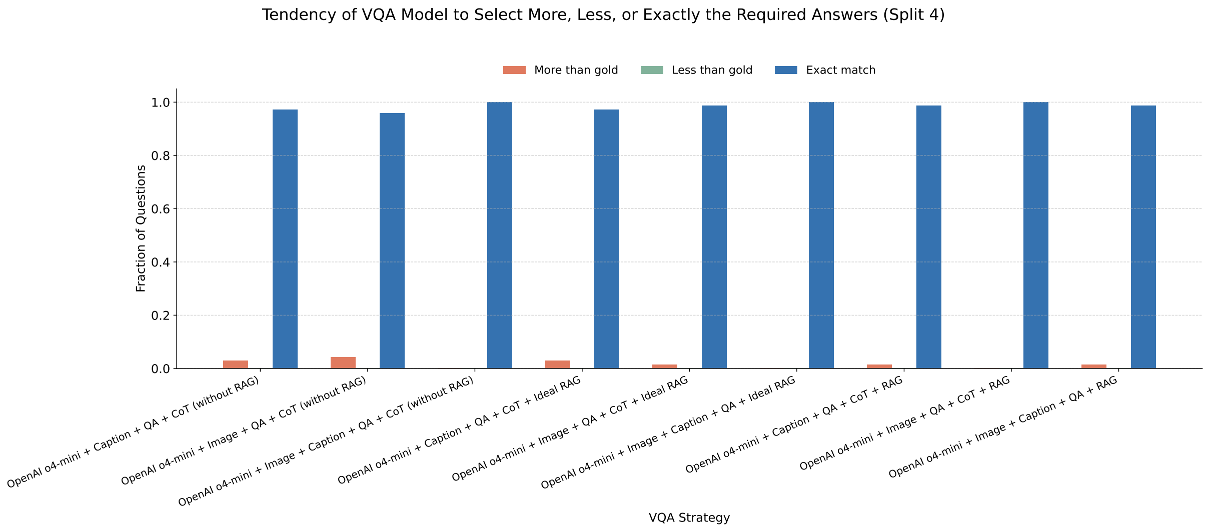


Figure 48: Tendency of OpenAI Models to Select More, Less, or Exactly the Required Answers (VQA, Split 4)

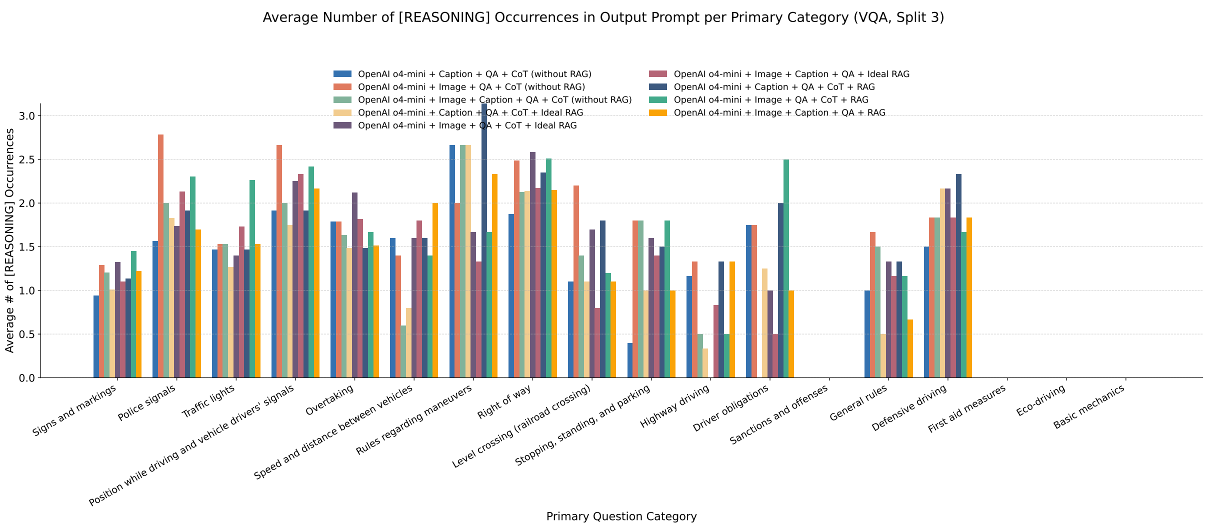


Figure 49: Average Number of reasoning steps in Output per Category (VQA, Split 3)

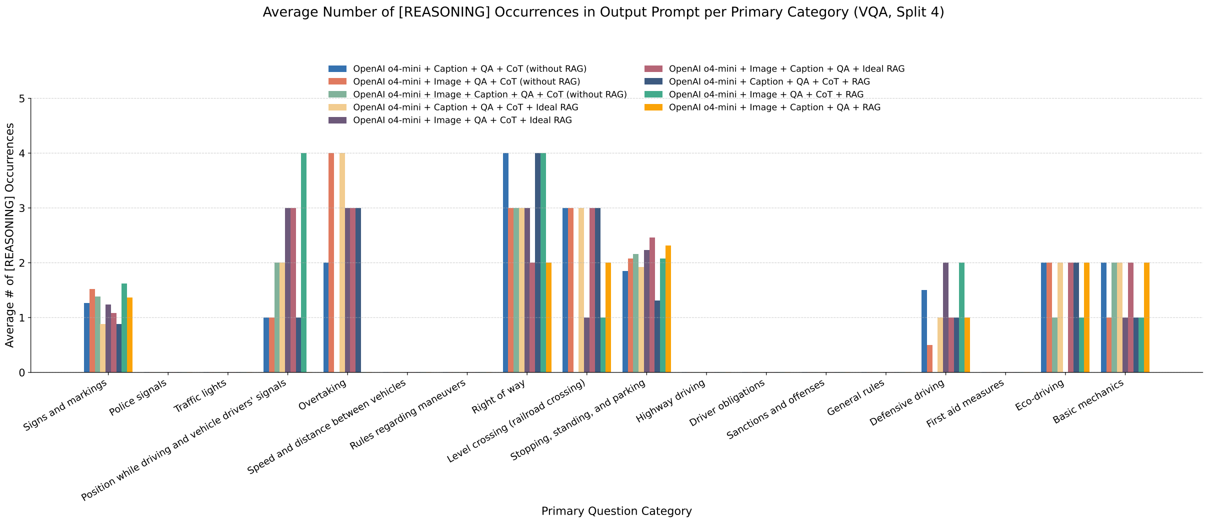


Figure 50: Average Number of reasoning steps in Output per Category (VQA, Split 4)

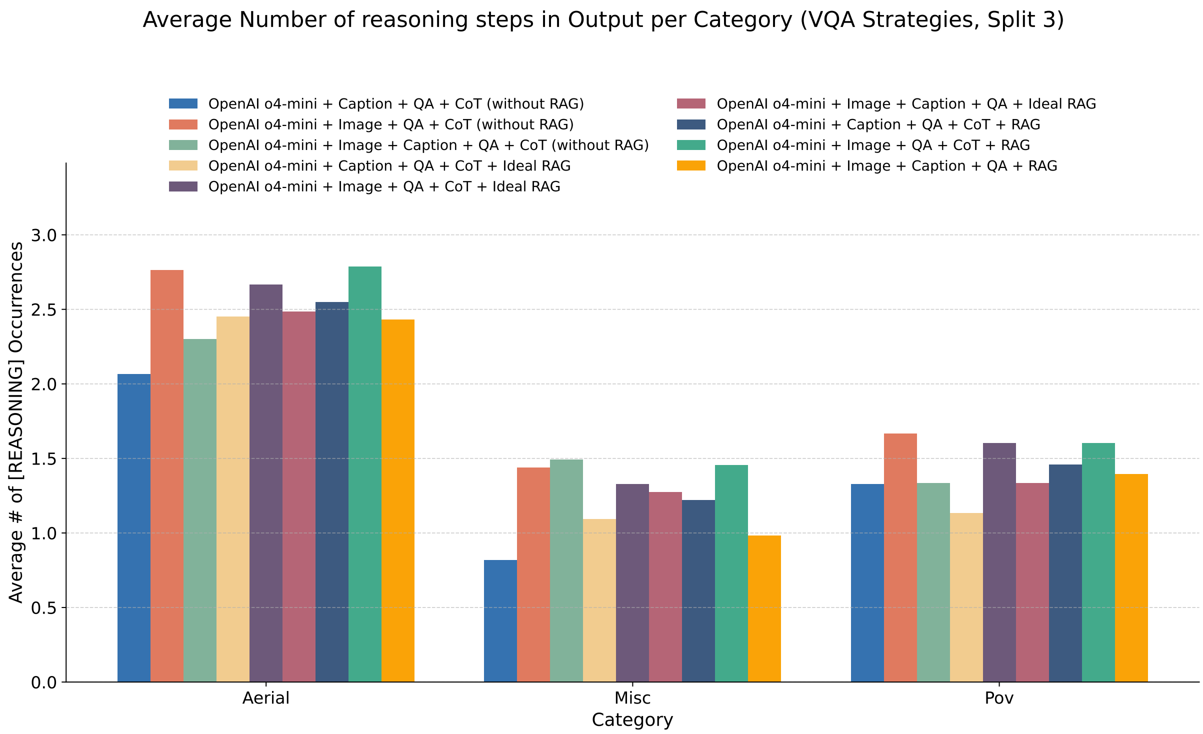


Figure 51: Average Number of reasoning steps in Output per Secondary Category (VQA, Split 3)

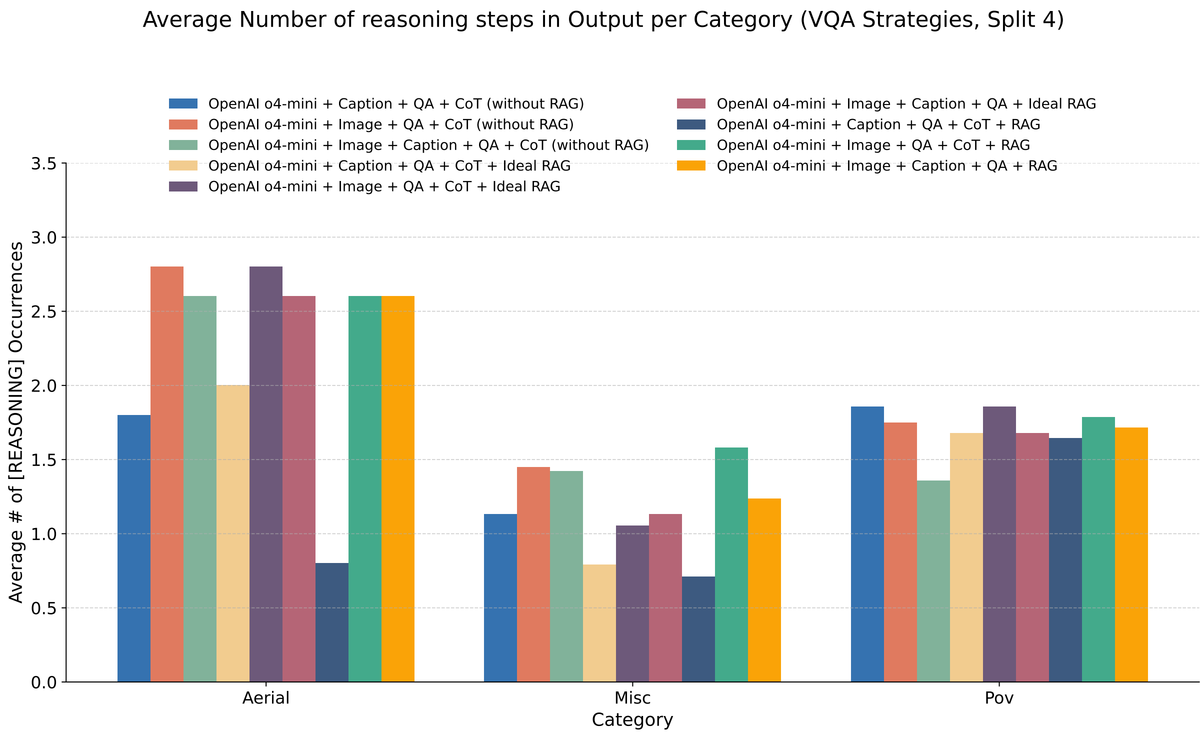


Figure 52: Average Number of reasoning steps in Output per Secondary Category (VQA, Split 4)