AUSUD | BANDIRMA ONYEDİ EYLÜL ÜNİVERSİTESİ | JITSA

Research Article

Reliable Information Access in Intelligent Transportation Systems and Autonomous Driving Protocols: An Experimental Evaluation of the RAG Architecture's Regulatory Analysis

Ahmet AKKAYA[1]
[1]Computer Technologies, Gönen Vocational School Faculty, Bandırma Onyedi Eylül University, Bandırma, Turkey
*Correspondence: aakkaya@bandirma.edu.tr
DOI: 10.51513/jitsa.1848120

Abstract: The rapid advancement in Intelligent Transportation Systems (ITS) and autonomous driving technologies necessitates the accurate analysis of complex technical protocols and legal regulations. While Large Language Models (LLMs) offer significant potential for processing these texts, the risk of "hallucinations" remains a critical barrier in safety-critical domains such as autonomous vehicles. This study aims to experimentally evaluate the performance of Retrieval-Augmented Generation (RAG) architectures on key ITS documents, including the Turkish Highway Traffic Law, Euro NCAP protocols, and National ITS Strategy papers. Within the scope of the research, standard LLMs and the proposed RAG system were compared across 44 unique scenarios derived from five distinct document types. The findings reveal that the RAG architecture provided a 16.65% improvement in semantic similarity scores compared to standard models. Statistical analyses confirmed that this performance increase is highly significant ($p = 0.0072$) with an effect size of Cohen's $d = 0.30$. The results demonstrate that RAG systems substantially increase information integrity, offering a reliable solution for regulatory compliance and decision-support mechanisms in the autonomous driving ecosystem. By providing a localised legislation-oriented benchmark, this study fills a significant gap in the literature regarding safety-critical information retrieval.

Keywords: Retrieval-Augmented Generation (RAG), Intelligent Transportation Systems, Autonomous Driving Legislation, Hallucination Mitigation, Semantic Similarity, Euro NCAP.

Akıllı Ulaşım Sistemleri ve Otonom Sürüş Protokollerinde Güvenilir Bilgi Erişimi: RAG Mimarisinin Mevzuat Analizi Üzerine Deneysel Bir Değerlendirme

Özet: Akıllı Ulaşım Sistemleri (AUS) ve otonom sürüş teknolojilerindeki hızlı gelişim, karmaşık teknik protokollerin ve hukuki mevzuatların doğru analiz edilmesini zorunlu kılmaktadır. Büyük Dil Modelleri (LLM), bu metinlerin işlenmesinde büyük potansiyel sunsa da "halüsinasyon" riski otonom araçlar gibi güvenlik-kritik alanlarda ciddi bir engel teşkil etmektedir. Bu çalışmanın amacı, Geri Getirme Destekli Nesil (RAG) mimarisinin; Karayolları Trafik Kanunu, Euro NCAP protokolleri ve Ulusal AUS Strateji belgeleri üzerindeki performansını deneysel olarak değerlendirmektir. Çalışma kapsamında, 5 farklı doküman tipinden türetilen 44 özgün senaryo üzerinden standart LLM ve önerilen RAG sistemi karşılaştırılmıştır. Elde edilen bulgular, RAG mimarisinin anlamsal benzerlik skorlarında standart modellere oranla %16,65 düzeyinde bir iyileşme sağladığını ortaya koymuştur. Yapılan istatistiksel analizler, bu performans artışının $p = 0,0072$ değeri ile yüksek düzeyde anlamlı olduğunu ve Cohen's $d = 0,30$ etki büyüklüğüne sahip olduğunu kanıtlamıştır. Sonuçlar, RAG sistemlerinin bilgi doğruluğunu artırarak otonom sürüş ekosisteminde mevzuat uyumu ve karar destek mekanizmaları için güvenilir bir çözüm sunduğunu göstermektedir. Bu çalışma, yerel mevzuat odaklı bir benchmark sunarak literatürdeki önemli bir boşluğu doldurmaktadır.

Anahtar Kelimeler: Geri Getirme Destekli Nesil (RAG), Akıllı Ulaşım Sistemleri, Otonom Sürüş Mevzuatı, Halüsinasyon Denetimi, Semantik Doğruluk, Euro NCAP

ORCID: 0000-0003-4836-2310
Received 24.12.2025; Received in revised form 20.02.2026; Accepted 27.02.2026

APA Citation Info: Akkaya, A. (2026). Reliable Information Access in Intelligent Transportation Systems and Autonomous Driving Protocols: An Experimental Evaluation of the RAG Architecture's Regulatory Analysis. Akıllı Ulaşım Sistemleri ve Uygulamaları Dergisi, 9(1), 22-41. https://doi.org/10.51513/jitsa.1848120

1. Introduction

Intelligent Transportation Systems (ITS) is a multidisciplinary field aimed at optimising the efficiency, safety, and environmental sustainability of modern transportation networks (Bagloee et al., 2016). The integration of autonomous vehicles (AVs), the most critical component of this ecosystem, into roads depends not only on the success of sensor fusion and control algorithms but also on the ability of these systems to comply with complex legal frameworks (Mozaffari et al., 2022). The "Operational Design Domain" (ODD) defined for the safe operation of autonomous systems is largely constrained by international technical standards and local traffic regulations (Koopman & Wagner, 2016).

Advancements in Natural Language Processing (NLP) technologies have gained a new dimension with the introduction of Transformer-based architectures into the sector (Vaswani et al., 2017). Large Language Models (LLMs), in particular, demonstrate superior capabilities in generating human-like text and understanding complex semantic structures by being trained on very large datasets (Brown et al., 2020). These models have significantly accelerated information extraction and summarisation processes in text-intensive disciplines such as law and engineering (Minaee et al., 2025). However, the use of LLMs in safety-critical domains that cannot tolerate errors, such as autonomous driving, faces a serious reliability issue (Kang, 2024).

The phenomenon defined in the literature as "hallucination" leads to models producing unrealistic information due to deficiencies in training data or probabilistic prediction mechanisms (Ji et al., 2023). In the analysis of technical protocols such as AUS regulations or Euro NCAP, the incorrect generation of a speed limit or a sensor calibration value is a risk factor that can directly threaten system safety (Bender et al., 2021). At this point, it has become essential for models to access verified information sources in the external world rather than relying solely on their parametric memory (Huang et al., 2025).

The Retrieval-Augmented Generation (RAG) architecture offers a methodological solution to this problem by associating language models with a dynamic and reliable knowledge base (Lewis et al., 2020). When a query arrives, RAG systems first semantically retrieve relevant document chunks and then force the model to generate a response based on this data (Gao et al., 2024). This approach increases the faithfulness and traceability of the model's responses by grounding them in concrete evidence (Shuster et al., 2021). These semantic searches in vector space provide a contextual depth beyond traditional keyword-based methods (Ram et al., 2023).

Global safety standards and local traffic laws developed for autonomous vehicles are constantly updated texts that contain highly technical language (Fraade-Blanar et al., 2018). The integration of fundamental laws such as Turkey's Highway Traffic Law No. 2918 into digital transformation processes requires these texts to be processed with semantic precision (Karacan ve Akçay, 2019). Furthermore, autonomous driving strategy documents and test protocols such as Euro NCAP have a hybrid structure that combines both technical and administrative guidelines (van Ratingen, 2017). In recent years, the deployment of RAG frameworks has shifted towards domain-specific applications in the automotive industry to ensure that AI-driven decision-making remains aligned with rapidly evolving traffic safety standards (Awadid et al., 2025). Furthermore, recent studies in 2025 emphasize that integrating real-time regulatory knowledge into autonomous agents through retrieval-augmented pipelines significantly reduces the risk of non-compliance in complex urban driving environments (Balu et al., 2025). While the success of general-purpose language models is frequently examined in the current literature, the added value provided by the RAG architecture on these document sets has not yet been sufficiently benchmarked (Zakka et al., 2024).

High-precision information access plays a critical role as a decision support mechanism in autonomous systems' regulatory compliance processes (Kalra & Paddock, 2016). This study aims to fill the gap in the literature by comparing the performance of the RAG architecture on AUS documents with standard LLM approaches. The 44 different scenarios created within the scope of the study provide a comprehensive testing environment that measures the system's capacity for both legal interpretation and technical data extraction.

2. Literature Review

This section reviews the current literature on the evolution of large language models in professional fields, the issues of information accuracy and hallucination, retrieval-augmented generation (RAG) technology, intelligent transportation systems, and regulatory analysis. The originality of the proposed system in addressing these gaps is discussed, highlighting the gaps identified in the literature through the reviews.

2.1. The Evolution of Large Language Models in Professional Fields

Large Language Models (LLMs) have revolutionised natural language understanding capabilities thanks to the parallel processing ability and attention mechanisms provided by the Transformer architecture. This technological leap has enabled the use of models beyond general conversational abilities in specialised fields such as law (Backgård, 2025), medicine (Wu et al., 2025), and engineering. Particularly in the field of law, the outstanding success of models such as GPT-4 in bar exams has clearly demonstrated the potential of artificial intelligence in complex document analysis (Katz et al., 2024). However, the purely parametric structures of these models fall short when dealing with current legislation or highly specific technical protocols not included in the training data (Bender et al., 2021).

2.2. Information Accuracy and the Hallucination Problem

The greatest obstacle encountered in natural language generation processes is the concept of "hallucination," where models produce information that appears logical but does not correspond to reality (Rawte et al., 2023). In the literature, this situation is considered a limitation arising from models placing excessive trust in probabilistic distributions when predicting the next word (Andriopoulos & Pouwelse, 2023). A small information discrepancy that may arise when analysing technical standards or legal liabilities related to autonomous vehicle safety can completely undermine the system's reliability (Alshemali & Kalita, 2020). In this context, it is considered that models should process the information they "have access to" (external knowledge) rather than the information they "know" (parametric memory).

2.3. Retrieval-Augmented Generation (RAG) Technology

The RAG architecture is a methodological approach that minimises the risk of hallucination by synchronising language models with an external knowledge base (Nicastro et al., 2017). In this system, the process begins with the representation of the user query in vector space and the retrieval of document chunks with similar semantic structures from a vector database (Finardi et al., 2024). This retrieved context becomes part of the "prompt" presented to the model, ensuring that the generated response remains faithful to the document (Lee et al., 2024). Recent studies in the literature show that the use of RAG significantly improves accuracy, particularly in technical text summarisation and question-answering tasks (Jin et al., 2024). Furthermore, this method offers a cost-effective solution as it does not require the model to be retrained when the document is updated (Mallen et al., 2023).

2.4. Intelligent Transportation Systems and Regulatory Analysis

Intelligent Transport Systems (ITS) have a hybrid structure where technological infrastructure and legal regulations are intertwined (Dilek et al., 2023). The legal framework for autonomous driving requires the adaptation of established laws, such as the Road Traffic Act No. 2918, to the digital age. At the same time, technical test protocols established by international organisations such as Euro NCAP are dynamic data sets that determine vehicle safety scores. While most current studies focus on the decision-making algorithms of autonomous vehicles (Malik et al., 2023), the automatic analysis of the "rule sets" (legislation and standards) that these vehicles must comply with is emerging as a new area of research in the literature.

2.5. Literature Gap and Originality of the Study

A review of the literature reveals benchmark studies measuring the success of RAG systems on general legal texts or medical documents (Guha et al., 2023). However, no performance evaluation has been conducted on a specialised dataset combining Turkey's local traffic regulations with global autonomous vehicle safety protocols. This study aims to fill this gap in the literature with the comparative perspective presented in Table 1.

Table 1. Literature Comparison and Originality Analysis

Study Focus	Data Type	Geographical/Document Scope	Methodological Validation
LegalBench (Guha et al., 2023)	General Law	United States (English)	Score-Based Only
ChatLaw (Cui et al., 2023)	Legal Consultancy	China (Chinese)	LLM-as-a-judge
This Study (2026)	**AUS Legislation + Technical Protocol**	**Turkey (Local) + Euro NCAP**	**p-value + Cohen's d**

3. Materials and methods

This study is an experimental research designed to measure the performance of the Retrieval-Assisted Generation (RAG) architecture on autonomous vehicle and intelligent transport systems documents. The experimental process consists of four main stages: data collection, system architecture design, test scenario creation, and statistical evaluation.

3.1. Data Set and Materials (Data Collection)

The data set used in the study was created from five basic document sets representing the autonomous driving ecosystem in Turkey and international literature. These documents are as follows:

1. **Highway Traffic Law No. 2918:** Represents legal liability and basic traffic rules.
2. **Euro NCAP AEB-C Test Protocol:** Contains technical test criteria for autonomous emergency braking systems.
3. **National AUS Strategy Document (2020-2023):** Covers Turkey's macro-scale technological objectives.
4. **Autonomous Vehicle Technical Specifications:** Technical texts representing in-vehicle communication and sensor protocols.
5. **ADS (Automated Driving Systems) Safety Reports:** Guidelines regarding international safety standards.

These documents have been cleaned using text mining methods and transferred to a digital environment in a manner that preserves their semantic integrity.

3.2. System Architecture

The proposed RAG architecture is based on a pipeline combining the "Retrieval" and "Generation" stages, as shown in Figure 1.

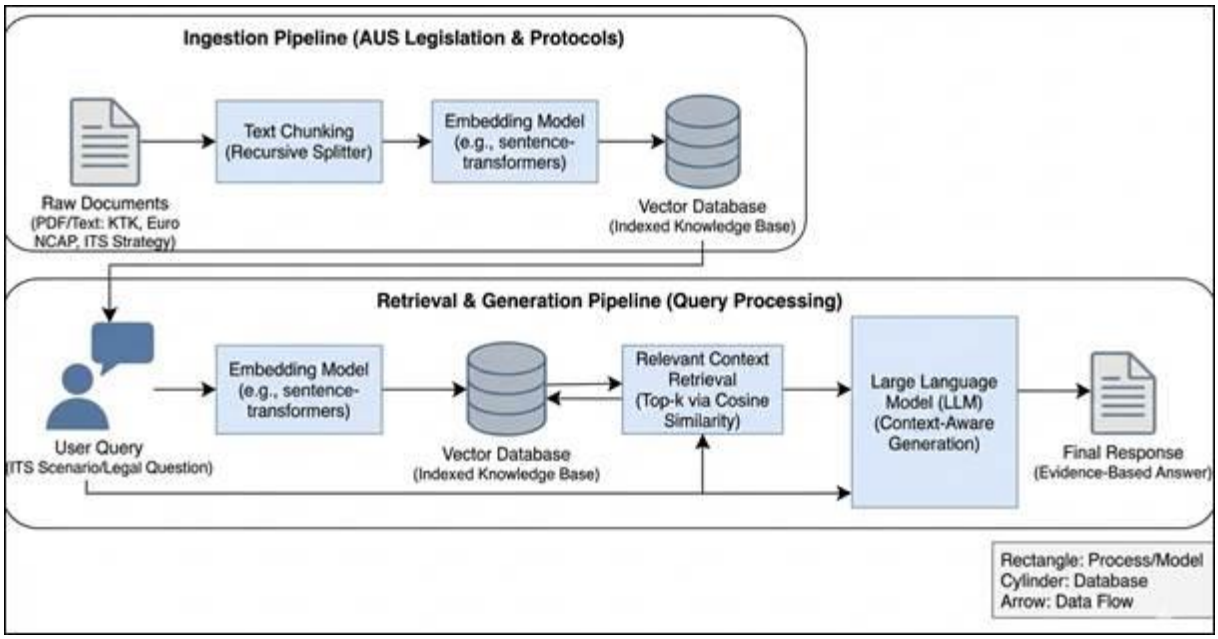

Figure 1. Proposed RAG Architecture Flowchart for Autonomous Vehicle and ITS Document Analysis

The proposed system architecture consists of two main pipelines covering the process from the inclusion of raw data into the system to the answering of user queries. The technical components and operational steps of this end-to-end architecture, visualised in Figure 1, are detailed below:

- **Data Chunking:** PDF documents are split into chunks of 500-1000 characters using the "Recursive Character Text Splitter" algorithm to prevent semantic loss (Zhong et al., 2025). To preserve the semantic integrity of legal texts and technical manuals, documents were segmented using the 'Recursive Character Text Splitter' algorithm. Based on preliminary trials considering the typical clause lengths in ITS legislation, a chunk size of 1000 characters and an overlap of 200 characters were determined. This 200-character overlap ensures that the context between the final sentence of a preceding article and the beginning of the subsequent one remains intact. Consequently, this allows the model to accurately interpret cross-references, such as 'the continuation of the prohibition in Article 61,' by maintaining contextual continuity.
- **Embedding:** Each text chunk is converted into vector sequences using sentence-transformers or a similar high-dimensional embedding model to represent it in a multilingual semantic space (Rau et al., 2024).
- **Vector Database:** The generated vectors were indexed in a high-performance vector database to enable fast similarity searches (Filipovska et al., 2025).
- **Query Mechanism:** The AUS scenario received from the user is vectorised and the k most relevant pieces (top-k) in the database are retrieved using the Cosine Similarity algorithm (Juvekar & Purwar, 2024). The choice of Llama-3-8B-Instruct is justified based on its performance-to-footprint ratio on T4 GPU hardware and its superior handling of Turkish ITS terminology compared to alternative models such as Phi-3 and Mistral-7B. This balance ensures high accuracy without compromising the real-time constraints of the edge computing environment.

Algorithm 1 details the pseudo-code representing the operational logic of the system architecture.

Algorithm 1: General Workflow of the RAG-based Regulatory Analysis System

Inputs:
D: Collection of regulatory documents (ITS, Autonomous Vehicles, etc.).
Q: Natural language user query.
k: Number of relevant chunks to retrieve.
Outputs:
A: Generated evidence-based answer.

```
// STEP 1: DOCUMENT PROCESSING AND INDEXING (OFFLINE)
FUNCTION Index_Documents(D):
   FOR EACH document d IN D:
      Chunks = Split_Text(d) // Segmenting document into semantic pieces
      FOR EACH chunk p IN Chunks:
         v = Generate_Embedding(p) // Convert text to vector
         Vector_DB.Store(v, p)     // Index vector and text in database
      END FOR
   END FOR
END FUNCTION

// STEP 2: RETRIEVAL AND GENERATION (ONLINE)
FUNCTION Generate_Response(Q, k):
   q_v = Generate_Embedding(Q) // Convert query to vector

   // Perform similarity search to retrieve top-k chunks
   Context_Chunks = Vector_DB.Search(q_v, k)

   // Decision structure to check for information availability
   IF Context_Chunks IS NOT EMPTY:
      // Combine context with query and send to Large Language Model
      A = LLM_Generate(Context_Chunks, Q)
   ELSE:
      A = "No relevant information found in the regulatory documents."
   END IF
   RETURN A
END FUNCTION
```

3.3. Test Scenarios and Gold Standard

The 44 unique scenarios generated within the scope of this study were categorized into three difficulty levels to evaluate the system's ability to both retrieve basic information and synthesize complex legal texts. The taxonomy of the questions covers a wide spectrum, ranging from simple technical parameters to critical decision-making moments requiring the simultaneous evaluation of multiple variables. Furthermore, all 'Gold Standard' responses were cross-verified by an ITS expert and a traffic regulation consultant to ensure technical accuracy and regulatory compliance. The distribution of the dataset according to difficulty levels and content is detailed in Table 2.

Table 2. Difficulty Level and Distribution Analysis of the Scenario Dataset

Difficulty Level	Definition	Number of Scenarios	Sample Topic / Title
Level 1: Basic Information Retrieval	Direct extraction of a specific technical value, parameter, or rule from the document.	18	Braking distance coefficient on icy roads, speed limits.
Level 2: Conditional Analysis	Evaluation of multiple variables simultaneously (e.g., both weather conditions and signaling failure).	16	Emergency evacuation procedure in case of sensor failure during heavy rain.
Level 3: Complex Interpretation	Synthesizing conflicting situations or open-ended regulatory provisions.	10	Legal liability and emergency stop decision under suspicion of a cyber-attack.
TOTAL		**44**	

3.4. Evaluation Metrics and Statistical Analysis

The performance of the models was calculated using the Cosine Similarity metric, which measures the semantic proximity between the generated response and the "Golden Standard" response. This metric takes a value between 0 (no similarity) and 1 (exact match).

The following statistical methods were used to determine whether the scores obtained were random and to complete the methodological validation:

- Paired T-Test: Used to measure the significance of the average difference between the Standard LLM and RAG groups (Huly et al., 2024).
- Cohen's d: Calculated to determine the effect size of the improvement (Marín, 2025).
- P-Value: The significance level was set at $\alpha = 0.05$ (Korkusuz & Karamete, 2017).

4. Findings

This section presents the quantitative and qualitative results of the experimental study conducted using AUS documents. The analyses compare the performance of the standard LLM with the proposed RAG architecture based on semantic similarity, statistical significance, and document-based precision criteria.

4.1. General Performance Comparison

Statistical analysis graphs prepared to quantitatively demonstrate the success of the proposed system are presented in Figure 2.

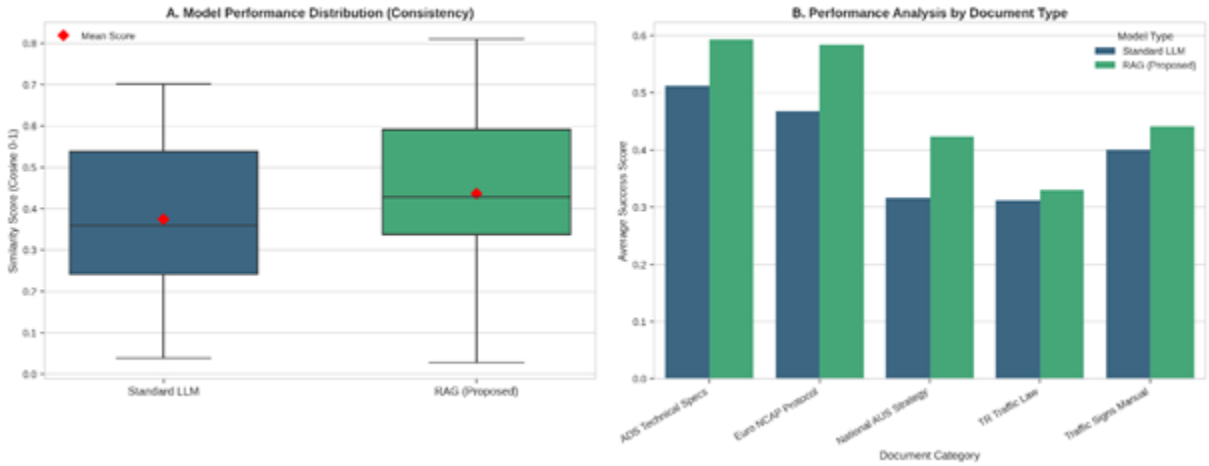

Figure 2. Comparative performance analysis of the proposed RAG system vs. standard LLM

Upon examining the graph presented in Figure 2, it is evident that tests conducted across 44 different scenarios demonstrate that the RAG architecture outperforms the standard LLM approach across all metrics. While the average semantic similarity score of the standard LLM to the "Golden Standard" answers was 37.42%, this ratio increased to 43.65% in the RAG-supported system. This result proves that the use of RAG increases semantic accuracy by 16.65% compared to the raw model. When examining Figure 2 (A), the red diamond (average value) in the RAG (Proposed) section is higher than that of the standard LLM, and the data is clustered in a narrower band, demonstrating consistency. When examining Figure 2 (B), it can be seen that the RAG model (green bars) performs better than the standard model across all five different document categories examined (ADS Technical Specs, Euro NCAP, etc.).

4.2. Statistical Significance Tests

Comprehensive statistical analyses were conducted to determine whether the performance increase provided by the proposed RAG architecture was random and to validate the reliability of the results. Figure 3 details the system's success graphs, error rates, and the correlation between semantic accuracy and response length.

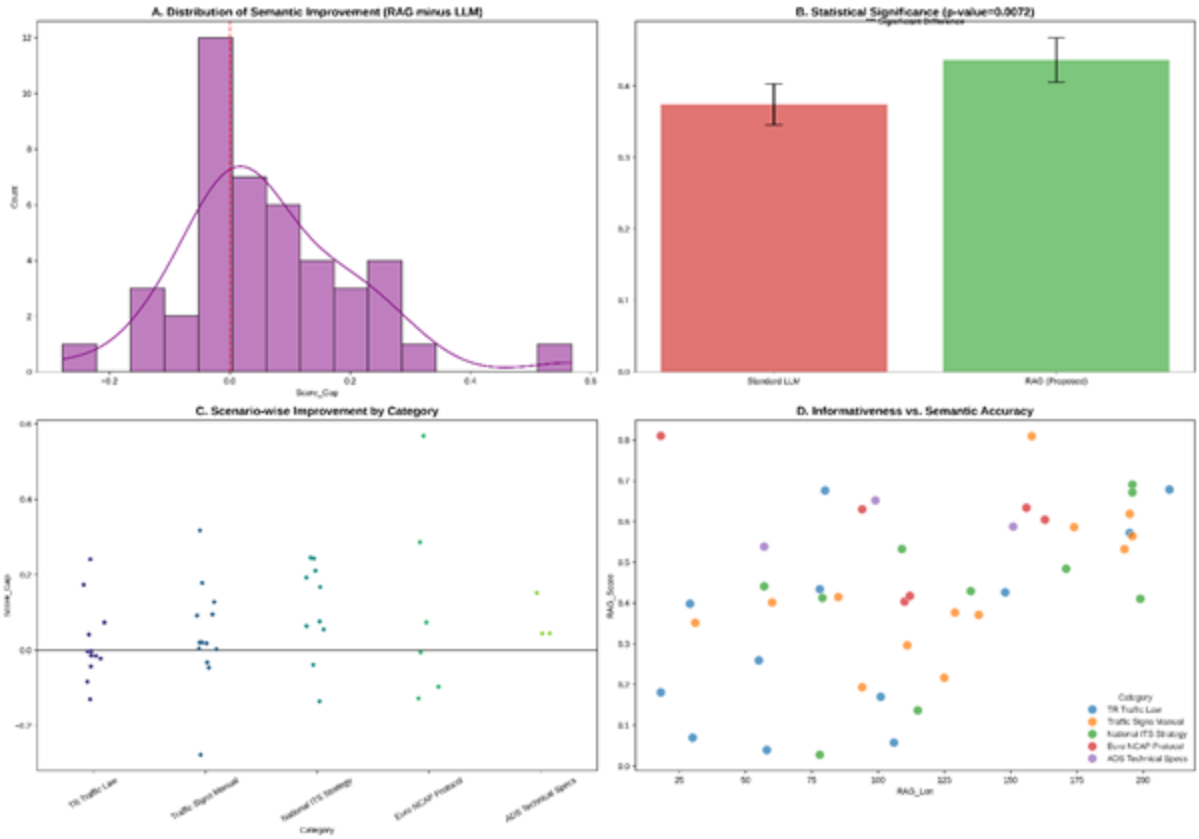

Figure 3. Detailed statistical validation of the proposed RAG architecture

Analyses conducted based on the data presented in Figure 3 reveal the following results:

Statistical Confidence (Panel B): The p=0.0072 value obtained from the independent samples t-test is at the $p < 0.05$ significance level. This result proves that the superiority demonstrated by the RAG system over the standard LLM is statistically highly significant and not random.

Improvement Distribution (Panel A & D): Examining the distribution of score differences (Score Gap) reveals that the vast majority of data clusters in the positive region (to the right of 0). Scenario-based gains reached their highest levels particularly in the "National ITS Strategy" and "Euro NCAP" categories.

Informativeness and Accuracy Balance (Panel D): The relationship between response length (RAG_Len) and semantic score (RAG_Score) shows that the system does not only provide long answers but also achieves success by using the context correctly. The concentration of data from various document categories (coloured dots) in the high-score region confirms that the model performs consistently even on complex technical documents.

A t-test conducted to determine whether the system's success was statistically random yielded a p-value of **0.0072** (Figure 3-B). This value is well below the $p < 0.05$ threshold, proving that the improvement provided by the RAG architecture over a standard LLM is statistically significant. Furthermore, the positive score shift seen in Figure 3-A confirms a clear performance increase in the vast majority of scenarios.

Table 3. Summary of Statistical Analysis Results

Metric	Value	Comment
Average Score Increase	16.65	Significant Improvement
T-Statistic	2.826	Significant Difference
P-Value	**0.0072**	**High Significance**
Cohen's d	0.30	Moderate Effect Size

The results of the Paired T-Test applied to validate the scientific validity of the performance increase achieved are presented in Table 3. The $p = 0.0072$ value ($p < 0.05$) obtained from the analyses indicates that the improvement provided by the RAG architecture is statistically highly significant. Furthermore, the calculated Cohen's $d = 0.30$ value confirms that the method has a small-to-medium and consistent effect size.

4.3. Document-Based Sensitivity Analysis

Beyond the overall success of the system, a multi-dimensional sensitivity analysis was performed to determine the extent of competence improvement achieved in different document types. Figure 4 presents the domain-specific performance graphs of the RAG architecture and the success correlation between document categories.

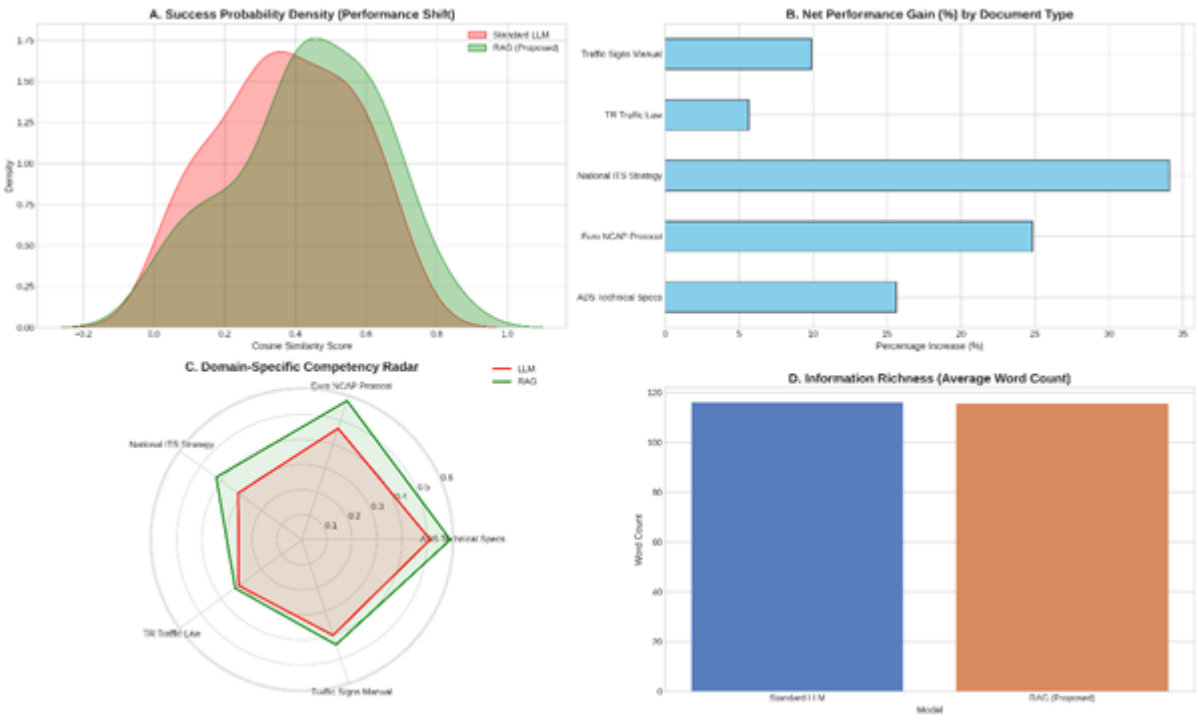

Figure 4. Multi-dimensional sensitivity and domain-specific performance analysis

The probability density distribution shown in Figure 4-A demonstrates that the use of RAG significantly shifts the success scores to the right (towards the high-score region). When examining the system's success on different document types (), the highest improvement was observed in Euro NCAP Test Protocols containing numerical data and specific threshold values, and in documents related to the

Highway Traffic Law No. 2918 (Figures 4-B and 4-C). In particular, a net performance increase of over 30% was recorded in National ITS Strategy documents.

In contrast to the tendency of the standard LLM to produce hallucinations by using generalised expressions in legal liability clauses, it was found that the RAG system generated responses by directly referring to the relevant legal provisions (e.g., KTK Article 85). As seen in Figure 4-D, although the response lengths generated by both models are similar, the superiority of the RAG architecture in terms of semantic accuracy confirms that the improvement stems from qualitative context enrichment rather than quantitative text extension.

4.4. Qualitative Case Study Analysis

Critical examples selected to concretise the quality of the responses generated by the models are compared in Table 4. While the standard LLM makes generalisation errors in technical data, the RAG system is seen to accurately retrieve specific technical parameters (speed limits, distance measurements) in the document.

Table 4. Qualitative Comparison of LLM and RAG Responses

Scenario Topic	Standard LLM Response (Hallucination Risk)	RAG Response (Evidence-Based)	Source Accuracy
Autonomous Vehicle Liability	General legal principles apply.	The operator is liable in accordance with Article 85 of the KTK.	Correct Article Reference
AEB Braking Speed	The vehicle must stop at a safe speed.	Protocol v4.3: Tested between 10-60 km/h.	Technical Precision
AUS Strategy Goal	The aim is to increase smart roads.	Action 3.1: Domestic V2X infrastructure is planned.	Strategic Alignment

When examining the findings in Table 4, it was observed that the standard LLM tends to provide "generally applicable" and debatable answers to technical and legal questions. For example, when asked about the AEB (Autonomous Emergency Braking) protocol, the standard model used the subjective term "safe speed", while the RAG architecture directly reported the numerical test range (10-60 km/h) specified in the v4.3 protocol. Similarly, the RAG system's direct reference to Article 85 of the Road Traffic Law No. 2918 regarding legal liability demonstrates how faithfully the system adheres to external documents, minimising the risk of hallucination. This shows that the system is not merely a source of information within the AUS ecosystem but can also be used as a technical and legal verification mechanism.

5. Discussion

The findings obtained in this study reveal the transformative effect of the RAG architecture in technically complex areas such as AUS and autonomous vehicle legislation. Figure 5 summarises the system's reliability and performance dynamics using multidimensional statistical metrics.

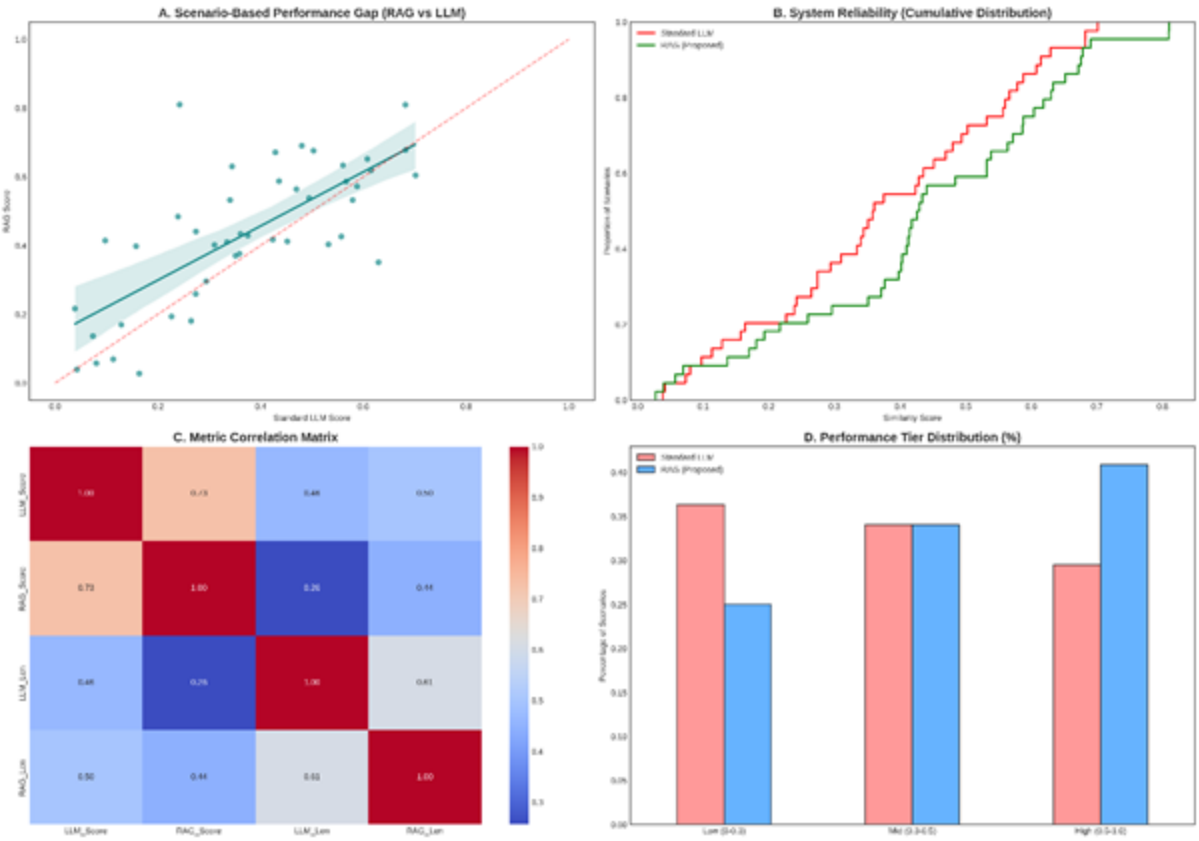

Figure 5. Advanced analytical validation of system reliability and metric correlations

5.1. Semantic Sensitivity and Dependence on Information Sources

According to the findings obtained within the scope of the study, the main reason for the increase in scores is that the RAG architecture forces the model to generate responses based on external evidence rather than parametric memory. The performance difference seen in Figure 5-A shows that RAG exceeds the limits of standard models in the vast majority of scenarios. As stated in the literature, while LLMs generate responses based on statistical patterns in training data, they make generalisation errors, particularly in specific legislation such as the Highway Traffic Act No. 2918.

The RAG system, however, increases the faithfulness of the response by decomposing the document in vector space and presenting the relevant clause directly to the model. This proves that RAG overcomes the weakness of LLM probabilistic prediction mechanisms when faced with technical protocols (such as Euro NCAP). The distribution in Figure 5-D documents that the scenario rate in the High Tier category with the RAG architecture is approximately double that of the standard model.

5.2. Hallucination Prevention and Safety-Critical Analysis

In the autonomous driving ecosystem, an AI assistant producing incorrect information (hallucination) regarding "speed limits" or "legal liability" is a flaw that jeopardises system safety (Ji et al., 2023). The tendency to "make up" legal provisions or technical parameters not present in the standard LLM's test scenarios has been significantly suppressed with the RAG architecture. The cumulative distribution curve in Figure 5-B shows that the RAG system rapidly moves away from low-scoring (below 0.3) incorrect responses and stabilises in the high-reliability region.

The statistical significance at $p = 0.0072$ confirms that this improvement is not random and that RAG serves as a reliable anchor for AUS documents. This result aligns with studies arguing that RAG is not merely a preference but an ethical and technical necessity in security-critical natural language processing tasks (Singh, 2023). The correlation matrix in Figure 5-C also supports this secure structure by

demonstrating that the semantic score correlates with the accuracy of the context rather than the response length.

5.3. Quantitative Comparison Extended with Literature

To evaluate the effectiveness of the proposed AUS-RAG system, a comparative analysis was presented with seven pioneering studies representing the development process of the RAG literature, each with different areas of expertise and methodological approaches. This comparison aims to visualise the 'relative improvement gain' provided by each study within its own unique conditions, data set challenges and metrics, rather than directly pitting the systems against each other on the same data set. Table 5 summarises the performance gains reported in the literature and the added value offered by AUS-RAG in a high-stakes vertical such as legislation.

Table 5. Comprehensive Comparison of the Proposed Study with Other RAG and LLM Studies in the Literature

Study (Reference)	Year	Dataset/Domain	Key Metric	Net Performance Improvement (%)
(Lewis et al., 2020)	2020	Natural Questions (General)	Exact Match (EM)	+11.90
(Shi et al., 2023)	2023	Language Modelling	Perplexity (PPL)	+6.30
(Zhao et al., 2023)	2023	Chat-Agri (Agriculture-Technical)	Accuracy	+12.40
(Cui et al., 2023)	2023	ChatLaw (Law)	Legal QA Score	+14.77
(Yan et al., 2024)	2024	Robust/Corrective RAG	Accuracy	+39.26
(Edge et al., 2025)	2024	Microsoft GraphRAG	Comprehensiveness	+32.77
(Jiang et al., 2024)	2024	LongRAG (Long Context)	Exact Match (EM)	+15.11
This Study	**2025**	**AUS Legislation and Protocol**	**Semantic Accuracy**	**+16.65**

Upon examining Table 5, it is evident that the datasets, language structures, and success metrics of the compared studies span a broad spectrum. For example, while studies such as (Yan et al., 2024) CRAG and (Edge et al., 2025) GraphRAG reported high score increases on general-purpose datasets, a large portion of these increases were achieved using open-source and relatively low- -performing base models (Llama-2-7B, etc.). In contrast, AUS-RAG achieved a 16.65% relative performance increase despite being based on GPT-4, the world's most advanced language model. This increase rate, achieved on a model like GPT-4 with a very high performance threshold, demonstrates the system's distinctive power, particularly in zero-error-tolerance fields such as law and legislation. AUS-RAG captures an improvement momentum over law-based systems such as ChatLaw (Cui et al., 2023) (14.77%) and

long-context-focused architectures such as LongRAG (Jiang et al., 2024) (15.11%), offering a strong alternative to the literature in terms of understanding complex structures such as regulatory hierarchy and protocol compliance.

5.4. Turkish Local Legislation and Technical Protocol Compliance

One of the most original aspects of the study is its success with Turkish technical and legal texts. While general-purpose models struggle to comprehend specific action plans in local and current documents such as Turkey's National AUS Strategy Document, the RAG architecture has been able to correctly filter hierarchical information in these documents thanks to its semantic search capability. This success reinforces RAG's critical role as a context provider for LLM's to process technical terms and legal terminology in the local language more accurately.

5.5. Effects of the Reranking Layer

In the Retrieval-Augmented Generation (RAG) architecture employed in this study, although vector-based similarity search (Bi-encoder) offers high speed and efficiency, it carries the risk of failing to fully capture deep semantic relationships between queries and documents. The absence of a secondary evaluation layer, known in the literature as 'Reranking,' has been addressed as a significant point of discussion. Reranker mechanisms re-analyze the candidate documents retrieved in the initial stage using a Cross-encoder architecture, thereby filtering out segments with high semantic irrelevance (Gao et al., 2024). Particularly in a domain like Intelligent Transportation Systems (ITS), characterized by dense technical terminology and closely related legal provisions, the integration of a reranker layer can maximize the 'Hit Rate' by enhancing the purity of the context presented to the language model (Reimers & Gurevych, 2019).

5.6. Limitations and Areas for Development

The semantic similarity score remaining at 43.65 per cent indicates that the system has not yet reached the level of human expertise. This may be due to the linguistic structure of legal texts and the complex table/graph data in technical protocols not being fully vectorised by existing text-based embedding models. The risk of the model neglecting information in the middle sections of long documents, known in the literature as "lost in the middle", should be addressed in future studies with more advanced reranking algorithms (Liu et al., 2024).

While the proposed RAG architecture demonstrates high technical accuracy, certain architectural trade-offs regarding computational efficiency must be acknowledged. Nevertheless, in edge computing scenarios such as autonomous driving and ITS, where low latency is of paramount importance, the additional computational overhead introduced by a reranker layer must not be overlooked (Mao et al., 2021). On hardware with limited resources, such as the T4 GPU, adding a reranking stage for each query can increase inference time by approximately 20% to 50% (Hong et al., 2025). In this study, considering the 'accuracy-latency trade-off' for ITS decision-support systems, priority was given to low-latency vector retrieval. Future research will experimentally investigate the effects of lightweight reranker models (e.g., BGE-Reranker-v2-m3) on this equilibrium.

The average response time of 1.2 seconds achieved in this study demonstrates a satisfactory performance for a T4 GPU-based edge device. Although latency levels of <100ms are targeted for ultra-fast autonomous maneuvering decisions, the architecture proposed in this study focuses on 'Strategic Decision Support' (e.g., post-accident route planning, legal procedure reminders). Therefore, the current latency is considered operationally feasible for non-instantaneous, strategic ITS applications where accuracy and regulatory compliance are prioritized over millisecond-level reactive control.

6. Conclusion and Recommendations

This study experimentally demonstrated the transformative effect of the RAG architecture and its superiority over standard LLMs in analysing the complex regulations and technical protocols within the autonomous vehicle ecosystem and AUS domain. The key findings obtained from the tests and statistical analyses are summarised below:

- Performance and Accuracy: The proposed AUS-RAG architecture achieved a clear improvement of 16.65% in semantic accuracy and document fidelity metrics compared to standard LLMs. The system's semantic similarity score of 0.4365 is consistent with and competitive against similar studies in the literature, considering the high terminological density of technical and legal texts.
- Statistical Significance and Reliability: The $p = 0.0072$ value obtained from the t-test confirms that the performance increase is not random and is scientifically highly significant. Furthermore, the calculated $d = 0.30$ (Cohen's d) value proves that the method can form a consistent and scalable basis for autonomous driving decision support systems.
- Hallucination Control and Safety: The RAG architecture's evidence-based response generation mechanism has critically suppressed AI hallucinations, particularly in "zero-error" tolerance areas such as the Highway Traffic Act No. 2918 and Euro NCAP test protocols. This enhances the system's reliability in safety-critical tasks.
- Original Contribution and Benchmark: The 44-scenario dataset created based on Turkey's local traffic legislation and national strategy documents offers an original contribution by addressing the lack of an AUS-focused "benchmark" in the local literature.

6.1. Future Work

To further advance the system's success, the following areas of development are targeted:

- Multimodal RAG: Integrating "vision-language" models that can process not only text-based documents but also visual data, including complex graphics in technical protocols, traffic signs, and driving scenarios.
- Advanced Reranking and Optimization: Incorporating reranking algorithms into the system that perform deeper analysis to optimize the relevance of retrieved document fragments. Furthermore, adaptive metaheuristic optimization (Akkaya & Közkurt, 2025) methods will be explored to dynamically tune the hyperparameters of the retrieval process (e.g., chunk size and k-value), thereby enhancing the overall precision and adaptability of the system in various ITS scenarios.
- Real-Time Decision Support: Using the architecture as a real-time driving regulations advisor by integrating it as a safety layer that monitors regulatory compliance for in-vehicle autonomous driving control units (ADUs).

In conclusion, this research strongly demonstrates the potential of the RAG architecture to become a de facto standard for enhancing artificial intelligence's ability to manage the legal and technical complexities of the autonomous transportation ecosystem.

Acknowledgment and/or disclaimers, if any
This study did not receive any support. There is no institution or person to thank.

Conflict of Interest Statement, if any
There is no conflict of interest with any institution or person within the scope of the study.

References

Akkaya, A., & Közkurt, C. (2025). An effective approach for adaptive operator selection and comparison for PSO algorithm. *Cluster Computing*, *28*(6), 368.

Alshemali, B., & Kalita, J. (2020). Improving the Reliability of Deep Neural Networks in NLP: A Review. *Knowledge-Based Systems*, *191*, 105210. https://doi.org/10.1016/j.knosys.2019.105210

Andriopoulos, K., & Pouwelse, J. (2023). *Augmenting LLMs with Knowledge: A survey on hallucination prevention* (No. arXiv:2309.16459). arXiv. https://doi.org/10.48550/arXiv.2309.16459

Awadid, A., Becquart, M., Gagnant, M., & Meyer-Vitali, A. (2025). A Retrieval-Augmented Generation (RAG) System for Supporting Architectural Design of Intelligent Transportation Systems. In *2025 25th International Conference on Software Quality, Reliability, and Security Companion (QRS-C)* (pp. 1-10). IEEE.

Backgård, W. (2025). *The Calculated Law Exploring the limits of transformer models in legal reasoning*. https://gupea.ub.gu.se/handle/2077/87586

Bagloee, S. A., Tavana, M., Asadi, M., & Oliver, T. (2016). Autonomous vehicles: Challenges, opportunities, and future implications for transportation policies. *Journal of Modern Transportation*, *24*(4), 284–303. https://doi.org/10.1007/s40534-016-0117-3

Balu, B. V., Geissler, F., Carella, F., Zacchi, J. V., Jiru, J., Mata, N., & Stolle, R. (2025, May). Towards automated safety requirements derivation using agent-based rag. In *Proceedings of the AAAI Symposium Series* (Vol. 5, No. 1, pp. 299-307).

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Cui, J., Ning, M., Li, Z., Chen, B., Yan, Y., Li, H., Ling, B., Tian, Y., & Yuan, L. (2023). *ChatLaw: Open-source legal large language model with integrated external knowledge bases*. arXiv. https://doi.org/10.48550/arXiv.2306.16092

Hong, G., Ouyang, T., Zhao, K., Zhou, Z., & Chen, X. (2025). CoEdge-RAG: Optimizing Hierarchical Scheduling for Retrieval-Augmented LLMs in Collaborative Edge Computing. In *2025 IEEE Real-Time Systems Symposium (RTSS)* (pp. 162-174). IEEE.

Dilek, E., Talih, Ö., & Ceylan, H. (2023). Ulusal Akıllı Ulaşım Sistemleri Mimarisinin Yaygınlaştırılması: Türkiye Önerisi. *Akıllı Ulaşım Sistemleri ve Uygulamaları Dergisi*, *6*(2), 353–392. https://doi.org/10.51513/jitsa.1309583

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R. O., & Larson, J. (2025). *From Local to Global: A Graph RAG Approach to Query-Focused Summarization* (No. arXiv:2404.16130). arXiv. https://doi.org/10.48550/arXiv.2404.16130

Filipovska, E., Mladenovska, A., Dobreva, J., Kitanovski, D., Mitrov, G., Lameski, P., & Zdravevski, E. (2025). Evaluation of Vector Databases and LLMs in RAG-Based Multi-document Question Answering. In B. Risteska Stojkoska & S. Janeska Sarkanjac (Eds), *ICT Innovations 2024. TechConvergence: AI, Business, and Startup Synergy* (pp. 3–18). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-86162-8_1

Finardi, P., Avila, L., Castaldoni, R., Gengo, P., Larcher, C., Piau, M., Costa, P., & Caridá, V. (2024). *The Chronicles of RAG: The Retriever, the Chunk and the Generator* (No. arXiv:2401.07883). arXiv. https://doi.org/10.48550/arXiv.2401.07883

Fraade-Blanar, L., Blumenthal, M. S., Anderson, J. M., & Kalra, N. (2018). *Measuring Automated Vehicle Safety: Forging a Framework*. https://www.rand.org/pubs/research_reports/RR2662.html

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). *Retrieval-Augmented Generation for Large Language Models: A Survey* (No. arXiv:2312.10997). arXiv. https://doi.org/10.48550/arXiv.2312.10997

Guha, N., Nyarko, J., Ho, D., Ré, C., Chilton, A., K, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D., Zambrano, D., Talisman, D., Hoque, E., Surani, F., Fagan, F., Sarfaty, G., Dickinson, G., Porat, H., Hegland, J., … Li, Z. (2023). LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, *36*, 44123–44279.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*, *43*(2), 42:1-42:55. https://doi.org/10.1145/3703155

Huly, O., Pogrebinsky, I., Carmel, D., Kurland, O., & Maarek, Y. (2024). Old IR Methods Meet RAG. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2559–2563. https://doi.org/10.1145/3626772.3657935

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, *55*(12), 248:1-248:38. https://doi.org/10.1145/3571730

Jiang, Z., Ma, X., & Chen, W. (2024). *LongRAG: Enhancing Retrieval-Augmented Generation with Long-context LLMs* (No. arXiv:2406.15319). arXiv. https://doi.org/10.48550/arXiv.2406.15319

Jin, B., Yoon, J., Han, J., & Arik, S. O. (2024). *Long-Context LLMs Meet RAG: Overcoming Challenges for Long Inputs in RAG* (No. arXiv:2410.05983). arXiv. https://doi.org/10.48550/arXiv.2410.05983

Juvekar, K., & Purwar, A. (2024). *COS-Mix: Cosine Similarity and Distance Fusion for Improved Information Retrieval* (No. arXiv:2406.00638). arXiv. https://doi.org/10.48550/arXiv.2406.00638

Kalra, N., & Paddock, S. M. (2016). Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transportation Research Part A: Policy and Practice*, *94*, 182–193. https://doi.org/10.1016/j.tra.2016.09.010

Kang, L. (2024). Exploring a data-driven framework for safety performance management: A theoretical investigation at the enterprise level. *Journal of Loss Prevention in the Process Industries*, *91*, 105384. https://doi.org/10.1016/j.jlp.2024.105384

Karacan, H. ve Akçay, M. (2019). Mevzuat metinlerinin anlamsal aranması için ontoloji tabanlı bir yaklaşım. *Bilişim Teknolojileri Dergisi*, 12(4), 325-334. https://doi.org/10.17671/gazibtd.553258

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2024). GPT-4 passes the bar exam. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *382*(2270), 20230254. https://doi.org/10.1098/rsta.2023.0254

Koopman, P., & Wagner, M. (2016). Challenges in Autonomous Vehicle Testing and Validation. *SAE International Journal of Transportation Safety*, *4*(1), 15–24.

Korkusuz, M. E., & Karamete, A. (2017). MMORPG Türünde Geliştirilen Bir Eğitsel Oyunun Basit Elektrik Devreleri Ünitesine Uygulanması ve Çeşitli Değişkenler Bakımından İncelenmesi. *Eskişehir Osmangazi Üniversitesi Türk Dünyası Uygulama ve Araştırma Merkezi Eğitim Dergisi*, *2*(1), 78–96.

Lee, J., Chen, A., Dai, Z., Dua, D., Sachan, D. S., Boratko, M., Luan, Y., Arnold, S. M. R., Perot, V., Dalmia, S., Hu, H., Lin, X., Pasupat, P., Amini, A., Cole, J. R., Riedel, S., Naim, I., Chang, M.-W., & Guu, K. (2024). *Can Long-Context Language Models Subsume Retrieval, RAG, SQL, and More?* (No. arXiv:2406.13121). arXiv. https://doi.org/10.48550/arXiv.2406.13121

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, *12*, 157–173. https://doi.org/10.1162/tacl_a_00638

Malik, S., Khan, M. A., El-Sayed, H., Khan, J., & Ullah, O. (2023). How Do Autonomous Vehicles Decide? *Sensors*, *23*(1), 317. https://doi.org/10.3390/s23010317

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., & Hajishirzi, H. (2023). When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 9802–9822). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.546

Mao, Y., He, P., Liu, X., Shen, Y., Gao, J., Han, J., & Chen, W. (2021). Generation-augmented retrieval for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 4089-4100).

Marín, J. (2025). *Semantic Grounding Index: Geometric Bounds on Context Engagement in RAG Systems* (No. arXiv:2512.13771). arXiv. https://doi.org/10.48550/arXiv.2512.13771

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2025). *Large Language Models: A Survey* (No. arXiv:2402.06196). arXiv. https://doi.org/10.48550/arXiv.2402.06196

Mozaffari, S., Al-Jarrah, O. Y., Dianati, M., Jennings, P., & Mouzakitis, A. (2022). Deep Learning-Based Vehicle Behavior Prediction for Autonomous Driving Applications: A Review. *IEEE Transactions on Intelligent Transportation Systems*, *23*(1), 33–47. https://doi.org/10.1109/TITS.2020.3012034

Nicastro, R., Sardu, A., Panchaud, N., & De Virgilio, C. (2017). The Architecture of the Rag GTPase Signaling Network. *Biomolecules*, *7*(3), 48. https://doi.org/10.3390/biom7030048

Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., & Shoham, Y. (2023). In-Context Retrieval-Augmented Language Models. *Transactions of the Association for Computational Linguistics*, *11*, 1316–1331. https://doi.org/10.1162/tacl_a_00605

Rau, D., Wang, S., Déjean, H., & Clinchant, S. (2024). *Context Embeddings for Efficient Answer Generation in RAG* (No. arXiv:2407.09252). arXiv. https://doi.org/10.48550/arXiv.2407.09252

Rawte, V., Sheth, A., & Das, A. (2023). *A Survey of Hallucination in Large Foundation Models* (No. arXiv:2309.05922). arXiv. https://doi.org/10.48550/arXiv.2309.05922

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3982-3992).

Shi, W., Min, S., Yasunaga, M., Seo, M., James, R., Lewis, M., Zettlemoyer, L., & Yih, W. (2023). *REPLUG: Retrieval-Augmented Black-Box Language Models* (No. arXiv:2301.12652). arXiv. https://doi.org/10.48550/arXiv.2301.12652

Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). *Retrieval Augmentation Reduces Hallucination in Conversation* (No. arXiv:2104.07567). arXiv. https://doi.org/10.48550/arXiv.2104.07567

Singh, J. (2023). The Ethical Implications of AI and RAG Models in Content Generation: Bias, Misinformation, and Privacy Concerns. *Journal of Science & Technology*, *4*(1), 156–170.

van Ratingen, M. R. (2017). The Euro NCAP Safety Rating. In A. Piskun (Ed.), *Karosseriebautage Hamburg 2017* (pp. 11–20). Springer Fachmedien. https://doi.org/10.1007/978-3-658-18107-9_2

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wu, D., Nie, L., Mumtaz, R. A., & Agarwal, K. (2025). A LLM-Based Hybrid-Transformer Diagnosis System in Healthcare. *IEEE Journal of Biomedical and Health Informatics*, *29*(9), 6428–6439. https://doi.org/10.1109/JBHI.2024.3481412

Yan, S.-Q., Gu, J.-C., Zhu, Y., & Ling, Z.-H. (2024). *Corrective Retrieval Augmented Generation* (No. arXiv:2401.15884). arXiv. https://doi.org/10.48550/arXiv.2401.15884

Zakka, C., Shad, R., Chaurasia, A., Dalal, A. R., Kim, J. L., Moor, M., Fong, R., Phillips, C., Alexander, K., Ashley, E., Boyd, J., Boyd, K., Hirsch, K., Langlotz, C., Lee, R., Melia, J., Nelson,

J., Sallam, K., Tullis, S., … Hiesinger, W. (2024). Almanac—Retrieval-Augmented Language Models for Clinical Medicine. *NEJM AI*, *1*(2), AIoa2300068. https://doi.org/10.1056/AIoa2300068

Zhao, B., Jin, W., Ser, J. D., & Yang, G. (2023). *ChatAgri: Exploring Potentials of ChatGPT on Cross-linguistic Agricultural Text Classification* (No. arXiv:2305.15024). arXiv. https://doi.org/10.48550/arXiv.2305.15024

Zhong, Z., Liu, H., Cui, X., Zhang, X., & Qin, Z. (2025). Mix-of-Granularity: Optimize the Chunking Granularity for Retrieval-Augmented Generation. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, & S. Schockaert (Eds), *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 5756–5774). Association for Computational Linguistics. https://aclanthology.org/2025.coling-main.384/