

## SUPPORT FOR ROAD TRANSPORT MANAGEMENT – AI-BASED LEGAL QUESTION ANSWERING: A SEARCH-BASED APPROACH AND PILOT EVALUATION

Laleh DAVOODI<sup>1</sup>, Maryam TEIMOURI<sup>2</sup>, Iwona WASIELEWSKA-  
MARSZAŁKOWSKA<sup>3\*</sup>, Filip GINTER<sup>4</sup>

<sup>1</sup> University of Turku, Department of Computing, TurkuNLP; laleh.l.davoodi@utu.fi,  
ORCID: 0000-0002-4305-4274

<sup>2</sup> University of Turku, Department of Computing, TurkuNLP; mtebad@utu.fi, ORCID: 0009-0002-8316-7844

<sup>3</sup> Nicolaus Copernicus University in Toruń, Faculty of Economic Sciences and Management, Department of  
Logistics; iwasielowska@umk.pl, ORCID: 0000-0001-6830-7859

<sup>4</sup> University of Turku, Department of Computing, TurkuNLP; figint@utu.fi, ORCID: 0000-0002-5484-6103

\* Correspondence author

**Purpose:** Road logistics practitioners operate in a complex regulatory environment where compliance information is dispersed across numerous legal documents, making the identification and interpretation of relevant regulations time-consuming and challenging. This study evaluates an AI-based legal question-answering (QA) system that supports regulatory compliance through retrieval-augmented generation (RAG). The objective is to assess whether a lightweight retrieval-grounded approach can provide accurate, traceable, and operationally useful answers without requiring domain-specific model fine-tuning.

**Design/methodology/approach:** The study adopts a design science and experimental evaluation approach. A domain-specific RAG architecture was developed using a corpus of Polish and European road transport regulations comprising approximately 2365 pages of legal documentation. A dataset of 50 practitioner-oriented legal questions with verified answers and article-level references was constructed for evaluation. The system combines semantic retrieval using vector embeddings and FAISS indexing with large language model-based answer generation. Performance was assessed through retrieval accuracy, answer correctness, citation grounding, and human-in-the-loop prompt refinement experiments.

**Findings:** The results demonstrate that retrieval-grounded legal QA can effectively support compliance-related information retrieval in road logistics. Under one-shot conditions, the system achieved 59% answer accuracy and 26.3% fully grounded responses. With limited user interaction and prompt refinement, answer accuracy increased to 82% and fully grounded responses to 41%. Many failures were caused by query ambiguity rather than missing knowledge. The findings indicate that retrieval-grounded AI can provide transparent, document-linked responses suitable for operational decision support.

**Research limitations/implications:** The study is based on a pilot dataset of 50 questions and focuses on Polish and EU road transport regulations. Future research should investigate larger datasets, multilingual retrieval, and user-centered evaluations examining trust, usability, and adoption.

**Practical implications:** The proposed system can assist logistics managers, dispatchers, compliance officers, and transport operators in locating and interpreting legal requirements more efficiently. Citation-grounded answers improve transparency and support faster regulatory decision-making.

**Originality/value:** This study introduces one of the first retrieval-augmented legal QA systems specifically designed for road logistics compliance. It contributes a domain-specific Polish road transport QA dataset, a practical RAG-based compliance architecture, and empirical evidence on the feasibility of retrieval-grounded AI for operational regulatory support.

**Keywords:** Artificial intelligence; Road transport regulations; Road transport management, Retrieval-augmented generation; Compliance support.

**Category of the paper:** Research paper.

## 1. Introduction

Logistics plays a central role in supply chain management by creating place utility and ensuring that raw materials and finished goods move efficiently from suppliers to customers across regions and borders. It encompasses interconnected activities such as transportation, warehousing, inventory management, packaging, and information processing, with transport via road, rail, air, waterways, pipelines, and multimodal networks forming the backbone of most operations (Islam et al., 2013).

In the European Union, road freight transport remains the dominant mode, accounting for nearly 1.9 trillion tonne-kilometres annually and supporting the majority of domestic and cross-border trade flows (Eurostat, 2025). Given this scale, road logistics is critical to modern supply chains but is also highly exposed to regulatory, operational, and safety challenges. Similar issues are observed globally; for instance, the Australian road transport sector is characterized by fragmented supply chains, strong cost pressures, and poor safety outcomes, highlighting how complex subcontracting structures can undermine working conditions and complicate effective regulation (Rawling et al., 2017).

Harmonizing regulations across the supply chain is fundamental to building an efficient and sustainable national logistics system. Effective integration requires coordinated action among government, industry, and society to ensure consistent policies and aligned operations (Akbar et al., 2025). Within this environment, understanding legal and regulatory requirements becomes a core operational capability rather than a peripheral administrative task. Empirical evidence shows that compliance with road safety and transport laws enhances supply chain reliability, responsiveness, and agility by reducing accidents, disruptions, and delivery delays (Adu et al., 2023).

More broadly, transport logistics has been characterized as both an economic and legal institution, where freight flows are governed through contracts, regulatory standards, and formal legal frameworks. When these legal structures are weak or unclear, inefficiencies,

disputes, and coordination problems emerge, undermining system performance and sustainability (Filippova, Voronina, 2021). At the same time, many regulatory practices rely heavily on tacit, experience-based knowledge that is difficult to formalize, leading scholars to recommend expert involvement to complement technical analyses and ensure practical relevance (Glistau et al., 2024).

However, legal document understanding itself is challenging due to the growing volume, length, and linguistic complexity of legal texts. Regulations, contracts, and policies often contain specialized terminology and formal structures that require expert interpretation, making manual analysis time-consuming even for professionals. For logistics practitioners who must make rapid operational decisions, this complexity creates substantial cognitive and practical burdens. Consequently, automated approaches such as summarization, retrieval, and QA are increasingly necessary to support efficient and reliable decision-making (Quevedo, 2023).

Recent advances in artificial intelligence (AI), particularly large language models (LLMs), offer promising opportunities to address these limitations. Nevertheless, evidence suggests that general-purpose LLMs alone remain insufficient for legal question answering (QA), as they often generate overly generic responses, lack logical specificity, and may hallucinate information, potentially exposing users to legal and financial risks. Retrieval-augmented generation (RAG) approaches have therefore emerged to ground LLM outputs in relevant legal documents and cases, improving factual accuracy and reducing hallucinations (Yao et al., 2025). RAG is an AI framework that improves LLM outputs by retrieving relevant information from external knowledge sources to ground responses in accurate, up-to-date, and verifiable facts (Martineau, 2023).

Recent legal QA research demonstrates rapid progress in retrieval-augmented and LLM-based assistance across multiple jurisdictions and legal domains but also reveals recurring limitations. Several studies introduce domain and language-specific datasets and benchmarks, such as criminal-law QA in India with practitioner-informed evaluation (Nigam et al., 2023), and Thai legal QA benchmarks focusing on financial and tax law while systematically comparing RAG pipelines and long-context alternatives (Akarajadwong et al., 2025a, 2025b).

In China, LeDQA provides long case-document QA with a professional question schema and sentence-level relevance annotations (Liu et al., 2024). Other work targets longer, explanatory answers and interpretability, proposing reasoning-first generation and semi-synthetic long-form QA datasets (Ujwal et al., 2024), and showing that improving legal QA often requires more than semantic matching by integrating logical structures with retrieval and generation (Yao et al., 2025). Complementary Polish-language studies examine complex multi-step reasoning datasets (Wojtasik et al., 2025), retrieval from lay statements to statutory provisions and the effects of fine-tuning retrieval models (Smywiński-Pohl et al., 2025), and explainable clause retrieval over a civil-code knowledge base (Matak, Chudziak, 2025).

Finally, systematic evidence from broader Legal NLP literature emphasizes that legal QA, summarization, and retrieval remain data and evaluation-constrained, and that curated datasets and robust grounding are central bottlenecks (Quevedo et al., 2023).

However, despite these advances, none of the existing systems specifically address the operational needs of road logistics compliance. Therefore, across these studies, three gaps are especially relevant for road logistics. First, most benchmarks focus on general legal domains (civil/criminal/tax, procurement, case law) rather than operational transport compliance, where users need fast, actionable answers linked to specific regulations (Nigam et al., 2023; Akarajadwong et al., 2025a, 2025b; Liu et al., 2024; Karp et al., 2025).

Second, many evaluations remain offline and benchmark-driven (e.g., exam-style multiple choice or synthetic QA), and results repeatedly highlight persistent issues with hallucinations, incorrect citations, and brittle reasoning, especially for longer or judgment-style outputs, suggesting that system usefulness must be assessed in more assistive, user-facing settings (Karp et al., 2025; Ujwal et al., 2024; Yao et al., 2025).

Third, legal RAG performance is highly dependent on the availability of task-appropriate QA data; however, transport-specific datasets remain largely absent, particularly for road transport regulations and practitioner-oriented questions reflecting dispatcher or logisticsmanager decision contexts (Quevedo et al., 2023; Smywiński-Pohl et al., 2025).

Moreover, legal QA is inherently complex and data-intensive, while existing benchmark datasets are predominantly large, expert-annotated collections developed in English or Asian languages and focused on domains such as judicial decisions, contracts, or privacy law rather than operational transport regulations, creating a clear gap for Polish road-logistics compliance applications (Abdallah et al., 2023).

Leveraging these insights and addressing the identified gaps, we focus on road-transport compliance as an under-served but operationally critical setting for legal QA. We develop and evaluate a domain-specific, Polish-language, retrieval-grounded QA system that supports practitioners in locating and interpreting relevant provisions with transparent evidence links. To enable an evaluation aligned with realistic logistics decision-making, we curate a focused dataset of 50 road-transport legal questions and answers reflecting practitioner-style queries. Accordingly, we address the following research questions:

**RQ1.** *How can AI-based legal question answering be developed for Polish-language logistics and transport regulations, where no domain-specific systems currently exist?*

**RQ2.** *How accurately does the proposed RAG system retrieve the relevant legal basis and generate citation-grounded answers on a curated set of Polish road-transport compliance questions?*

Beyond addressing the research questions, this study contributes to the emerging intersection of logistics and legal AI by positioning legal QA within an operational compliance context rather than another legal QA benchmark. Specifically, we introduce a domain-specific

Polish road-transport QA dataset composed of practitioner-style, citation-grounded questions reflecting real logistics decision scenarios. Moreover, The originality of this study lies in its focus on road logistics compliance, a domain largely overlooked in existing legal QA research.

Our results provide empirical evidence that a lightweight retrieval-augmented architecture can support regulatory QA without domain-specific fine-tuning or large annotated corpora. Hence, the study reframes legal QA as an assistive decision-support tool for dispatchers and logistics managers and provides empirical insights into how linguistic specificity and jurisdictional cues influence retrieval behavior in multi-layered regulatory environments. Moreover, our results show that failures arise primarily from query formulation ambiguity rather than retrieval absence, indicating that interactive clarification is a structural requirement of regulatory QA systems.

The originality of the study lies in (1) developing a domain-specific RAG-based legal QA system for road logistics regulations, (2) constructing a practitioner-oriented Polish road transport QA dataset with verified legal references, and (3) empirically evaluating retrieval accuracy, answer correctness, and citation grounding in a real compliance-support setting. By positioning legal QA as an operational decision-support tool rather than solely a benchmark task, the study extends current research on legal AI and logistics compliance.

The rest of the paper is organized as follows: The *Literature Reviews* section presents related work on regulatory complexity in road logistics and retrieval-augmented legal QA. The *Methodology and Experimental Setup* section presents the methodology and experimental setup, including dataset construction, the RAG system architecture, and the evaluation framework and results. The *Discussion* section discusses the findings, and the *Conclusion* section concludes the paper. Finally, the *Future Direction* section outlines directions for future work.

## 2. Literature Review

Logistics and road transport operations increasingly function at the intersection of operational efficiency, regulatory compliance, and digital transformation. As supply chains become more global and interconnected, firms must simultaneously manage complex transport processes, evolve legal requirements, and adopt new digital technologies.

These developments have elevated regulatory knowledge and information management from peripheral administrative tasks to core operational capabilities. At the same time, advances in AI and data-driven systems are reshaping how logistics decisions are supported and executed. Against this backdrop, existing research spans three interrelated streams: regulatory and legal complexity in logistics practice, AI and retrieval-based approaches for legal QA, and technology acceptance factors influencing the adoption of digital decision-support tools.

Reviewing these streams together provides the foundation for understanding why intelligent, assistive systems are increasingly necessary to support compliance and decision-making in road logistics.

## **2.1. Regulatory Complexity and Decision-Making Challenges in Road Logistics**

Globalization and increasingly complex supply chains have made logistics and customs management central to the efficiency and compliance of international transport operations. While digital technologies such as automation, blockchain, and AI improve operational performance, their integration poses new challenges in meeting evolving regulatory and standardization requirements (Orłowska, Chackiewicz, 2024).

Social legislation in road transport seeks to protect drivers' working conditions and improve safety, yet EU and AETR rules are often perceived as complex and difficult to interpret. Compared with other countries, these regulatory burdens can hinder compliance and may even negatively affect road safety, highlighting the need for clearer and more supportive decision tools (Poliak et al., 2018).

Moreover, road logistics operates within a complex, multi-layered regulatory environment where compliance is not optional but foundational to effective operations. Evidence shows that companies cannot function efficiently without being aware of and integrating legal, governmental, and transport regulations into daily practices. Transportation and logistics laws directly shape supply chain performance, costs, and reliability, making regulatory knowledge a core operational competence rather than a peripheral requirement (Mutangili, 2019).

Similarly, previous research emphasize that legal transport operations are also governed by formal documentation requirements and evidentiary rules defined under national law and international conventions such as Convention on the Contract for the International Carriage of Goods by Road (CMR) and Uniform Rules Concerning the Contract of International Carriage of Goods by Rail (CIM) (Garnowski, 2017).

Beyond structural complexity, compliance challenges translate into tangible operational risks. Legal misunderstandings may lead to inspections, fines, shipment delays, and contractual liabilities. Sector-specific legal frameworks, such as national commercial goods transport law, further demonstrate the formal and codified legal structures governing freight, forwarding, and warehousing activities under domestic and international provisions (Walther, 2008).

It should also be noted that both EU and Polish regulations do not provide for uniform legal regulations covering all modes of transport. In the case of Polish legal regulations, the most general regulations concerning the transport of persons and goods are contained in the Civil Code. The provisions of the Civil Code in this regard (Odzimek, 2023). Consequently, day-to-day logistics decision-making frequently depends not on formal legal analysis but on tacit, experience-based knowledge accumulated by dispatchers and managers.

In addition, educational perspectives emphasize that legal literacy is a necessary competence for logistics professionals, as daily activities rely heavily on contracts, liability rules, and international regulations such as CMR and INCOTERMS (Odzimek, 2023). However, in practice, operators often work under time pressure and rely on experience-based judgments rather than systematic consultation of legal texts, which may increase the risk of non-compliance.

Recent developments in digital technologies suggest potential solutions to operational complexity in logistics. Bibliometric evidence indicates growing interest in the use of AI for decision-making support, process automation, and data-driven optimization in supply chains (Marchewka, Cierzniewska, 2025).

Leveraging these developments, AI-based tools may also assist practitioners in navigating complex regulatory information by improving retrieval, interpretation, and transparency of legal knowledge. Additionally, the literature suggests that regulatory complexity, operational risk, and tacit knowledge gaps collectively justify the development of AI-assisted legal QA systems tailored specifically to logistics practice.

## **2.2. Retrieval-Augmented Generation for Legal Question Answering**

AI is increasingly reshaping legal practice, with natural language processing (NLP) and LLMs supporting tasks such as legal document analysis, contract review, case prediction, legal research, and regulatory compliance. Recent studies report rapid growth in both the number and sophistication of these applications, highlighting the expanding role of AI as a practical decision-support tool for legal professionals (Siino et al., 2025). Despite enabling powerful QA capabilities, recent LLMs remain prone to hallucinations and factual errors (Zhai, 2025).

Fine-tuning a language model on legal corpora can improve stylistic alignment but does not guarantee up-to-date knowledge and requires costly retraining whenever regulations change. Moreover, fine-tuned models internalize knowledge parametrically, making source attribution and traceability difficult, which is an important limitation in compliance and sensitive contexts. To address these limitations, Lewis et al. (2020) introduced RAG, which combines language models with document retrieval to ground responses in external knowledge sources, thereby improving factual accuracy and transparency. RAG extends LLMs by combining text generation with external information retrieval. Instead of relying only on knowledge learned during training, the system retrieves relevant document fragments from a document database and incorporates them into the prompt used for answer generation. This process grounds responses in verifiable sources, improving factual accuracy, and reducing hallucinations. By separating knowledge retrieval from language modeling, RAG thus enables domain adaptation without requiring model fine-tuning or large annotated datasets (Hindi et al., 2025). These properties make RAG particularly well-suited for evolving legal and regulatory environments where precision, traceability, and adaptability are critical.

Legal QA ranges from simple binary or multiple-choice tasks to long-form explanatory responses and is commonly implemented as a two-stage pipeline that integrates semantic retrieval with LLM-based answer generation (Redelaar et al., 2024).

Building on this paradigm, recent studies have applied RAG-based legal QA across several national contexts. For instance, Matak and Chudziak (2025) develop a system based on the Polish Civil Code, contributing valuable evidence for non-English legal settings. Their work examines how chunking strategies influence retrieval efficiency and downstream performance. However, evaluation relies on an exam-style multiple-choice format in which the model selects among predefined answers, making it unclear whether the setup truly measures retrieval quality. As a result, answer selection may be conflated with genuine retrieval effectiveness.

Wiratunga et al. (2024) introduce a case-based reasoning approach for RAG using an Australian law QA dataset and relies on BERT-based models that are explicitly trained for retrieval and similarity matching. While methodologically sound, it shares a similar limitation with the first work: neither paper evaluates whether the retrieved information is actually useful in a human-in-the-loop setting. In particular, there is no assessment of how retrieval supports a user's learning or reasoning process, with the AI acting as an assistant rather than a standalone answer generator.

Prior work on Thai legal QA has largely focused on statutory domains such as financial law and tax law. These studies construct datasets and benchmarks exclusively from Thai legal documents, without incorporating English or other languages. The resulting QA datasets are verified against authoritative legal references to ensure factual correctness. However, these systems are evaluated primarily through offline benchmarking and automated metrics. They are not deployed or assessed in real-world usage scenarios, and no human users are involved in evaluating system performance in practice. Moreover, the systems are not designed or studied as tools for user learning, such as supporting legal education or practitioner knowledge acquisition, but rather as static benchmarks for model comparison (Akarajaradwong et al., 2025a, 2025b).

In the Indian legal domain, existing work has focused mainly on criminal law and Supreme Court judgments. These studies construct a primary dataset along with a smaller test set to evaluate QA over judicial texts. While QA pairs are provided, the datasets generally lack explicit citation grounding or reference validation, which limits their applicability in learning-oriented or trust-critical settings. Evaluation is conducted without involving human users interacting with the system in realistic legal scenarios. Additionally, these systems are not framed or evaluated as learning tools for users, such as law students or legal professionals, and remain monolingual with no multilingual evaluation (Nigam et al., 2023).

For Chinese civil law, recent efforts introduce QA datasets derived from legal case documents with limited human supervision. In this setting, large language models are used to annotate data under human oversight, but the task formulation is restricted to binary or ternary decision-making (e.g., “yes” versus “no or unknown”). As a result, these systems do not support open-ended answer generation or explanatory responses that would facilitate user learning. Evaluation focuses on annotation quality rather than on human interaction in real legal usage scenarios. Similar to the Thai and Indian datasets, the systems are monolingual, exclude English, and are not evaluated as educational or decision-support tools for end users (Liu et al., 2024).

Text embeddings convert raw text into numerical vector representations that allow machine learning systems to compute semantic similarity and carry out the relevant document retrieval in RAG. Earlier approaches such as TF-IDF rely on sparse frequency-based representations, while static embeddings like GloVe capture global word co-occurrence patterns (Pennington et al., 2014). More recent transformer-based models, including BERT, produce contextual embeddings that encode meaning based on surrounding text, enabling richer semantic understanding (Devlin, 2019; Muchori, Maina, 2025). To compare these embeddings, we use cosine similarity, which measures semantic relatedness by computing the angle between vectors and is widely adopted for retrieval and ranking tasks in embedding-based NLP systems (de Vos et al., 2021).

Taken together, the reviewed literature highlights that many legal QA systems are primarily evaluated using benchmark-style metrics, while comparatively little attention is given to how retrieved information supports practical reasoning and decision-making in applied contexts. In operational domains such as logistics compliance, however, the usefulness of AI systems depends not only on answer accuracy but also on the transparency and traceability of the supporting legal sources.

Accordingly, the architecture adopted in this study emphasizes practical interpretability in addition to technical retrieval performance. Rather than focusing solely on automated QA accuracy, the system is designed to expose the retrieval process and supporting evidence. Through a user-facing interface, practitioners can inspect the source document, page reference, and exact text segment used during answer generation. This design enables users to verify factual accuracy, assess relevance, and consult the broader legal context when necessary. In this way, the proposed RAG framework is positioned as an assistive decision-support tool that complements human interpretation rather than functioning as a fully automated black-box answer generator.

### 3. Methods

Prior research on legal QA often depends on large, annotated datasets and task-specific fine-tuning. In specialized operational domains such as road logistics, comparable QA datasets are scarce and building them at scale would require extensive manual effort. To address this constraint, we adopt a lightweight strategy based on a general RAG pipeline combined with targeted prompt engineering. The system grounds answers in a curated corpus of Polish and EU road-transport legal documents by retrieving relevant passages at query time and generating citation-linked responses, avoiding the need for training a domain-specific model on thousands of labeled question-answer pair examples.

We evaluate the feasibility of this approach using an offline test set of 50 Polish practitioner-style questions with verified answers and article-level references. Evaluation focuses on (i) retrieval success (whether the system retrieves the correct legal basis), (ii) answer correctness relative to the reference answers, and (iii) citation grounding quality (whether cited passages support the generated answer). This design enables a reproducible assessment of a domain-specific legal QA system in road logistics without relying on large-scale annotated corpora or user studies in the conference version.

#### 3.1. Dataset Construction and Data Preparation

Despite rapid progress in legal QA, there is currently no domain-specific dataset or evaluation framework targeting Polish road transport compliance. Developing such a dataset at scale would require substantial expert involvement to formulate practitioner-style questions, verify legally correct answers, and annotate precise citation spans. This creates a significant practical barrier to deploying AI-based compliance tools in logistics practice.

More broadly, recently developed open-domain question answering (OpenQA) systems typically depend on large-scale annotated datasets to achieve state-of-the-art performance. However, producing these resources is labor-intensive and costly, as it requires expert validation, span-level annotation, and citation grounding. As a result, high-quality legal QA benchmarks are available only for a limited number of languages, primarily English and Chinese (Rybak et al., 2024).

As summarized in Table 1, most widely used legal QA benchmarks focus on general legal domains such as case law, contract review, mergers and acquisitions, privacy policies, or national judicial examinations. These large datasets are predominantly in English or Chinese and are designed for broad legal reasoning tasks rather than sector-specific operational contexts.

**Table 1.**  
*Benchmark Legal Question Answering Datasets*

| Dataset                         | Legal Domain                                                                    | Size                                                                          | Language | Authors                  |
|---------------------------------|---------------------------------------------------------------------------------|-------------------------------------------------------------------------------|----------|--------------------------|
| CaseHOLD                        | U.S. case law (holding identification from citations)                           | 53,000+ multiple-choice questions                                             | English  | (Zheng et al., 2021)     |
| COLIEE 2024 - Task 1 and Task 2 | Federal Court of Canada case law                                                | 5000+                                                                         | English  | COLIEE 2024              |
| CUAD                            | Contract review / clause extraction                                             | 13,000                                                                        | English  | (Hendrycks et al., 2021) |
| MAUD                            | Merger Agreement Understanding                                                  | over 39,000 examples and over 47,000 total annotations                        | English  | (Wang et al., 2023)      |
| JEC-QA                          | National Judicial Examination of China                                          | 26,365 multiple-choice questions                                              | Chinese  | (Zhong et al., 2020)     |
| PolicyQA                        | privacy practices                                                               | 25,017 reading comprehension style examples and 714 human-annotated questions | English  | (Ahmad et al., 2020)     |
| CJRC                            | Chinese judicial reading comprehension                                          | 50,000 questions with answers                                                 | Chinese  | (Duan et al., 2019)      |
| PolQA                           | Criminal & civil court judgments; tax interpretations; administrative decisions | 7,000 questions, 87,525 manually labeled evidence passages                    | Polish   | (Rybak et al., 2024)     |

Source: Own materials.

Even recent retrieval-focused benchmarks such as LegalBench-RAG (Pipitone, Alami, 2024) evaluate retrieval performance across diverse legal corpora but do not target transport-specific regulations or compliance-oriented practitioner questions. Similarly, the Polish PolQA dataset (Rybak et al., 2024) provides an important contribution for Polish-language QA; however, it covers criminal and civil court judgments rather than operational road-transport regulations.

Importantly, none of these benchmarks are tailored to the domain of road transport compliance, where practitioners require fast, actionable, and citation-grounded answers linked to specific EU and national transport regulations (e.g., AETR, Regulation 561/2006, CMR, ADR). Developing a large-scale, fully annotated QA dataset for this specialized domain in Polish would require substantial expert effort to formulate realistic practitioner-style questions, verify answers, and annotate precise legal references. Given the dispersed and frequently updated nature of transport regulations, maintaining such a dataset would further increase the resource burden.

In order to evaluate our proposed approach, we constructed a domain-specific dataset of 50 Polish legal QA pairs covering core road-transport compliance topics (e.g., driving and rest times, cabotage, carrier liability, dangerous goods, and customs transit).

**Table 2.***Overview of the legal Q&A dataset fields with illustrative examples*

| Question / English                                            | Answer / English                                                                                                 | Question / Polish                        | Answer / Polish                                                                                                   | Document                                                                                                             | Cite                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | Page             |
|---------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------|------------------------------------------|-------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| What are the objectives of the legal regulations in the AETR? | The objective of the regulations included in the AETR is to improve road safety and drivers' working conditions. | Jakie są cele regulacji prawnych w AETR? | Odpowiedź: Celem regulacji ujętych w AETR jest poprawa bezpieczeństwa ruchu drogowego i warunków pracy kierowców. | European Agreement concerning the work of crews of vehicles engaged in international road transport (AETR). Preamble | BEING DESIROUS of promoting the development and improvement of the international transport of passengers and goods by road, CONVINCED OF the need to increase the safety of road traffic, to make regulations governing certain conditions of employment in international road transport in accordance with the principles of the International Labour Organization, and joined to adopt certain measures to ensure the observance of those regulations, HAVE AGREED | Page: No L 95/ 1 |

Source: Own materials.

The questions were developed based on a carefully selected set of 14 Polish and EU legal instruments governing road transport operations. The underlying corpus comprises legal materials in both English and Polish and covers key regulatory domains relevant to logistics and transport, including ADR requirements, the CMR Convention, customs procedures under the A.T.A. Carnet, and the TIR Convention.

In total, the corpus contains approximately 1,432,435 tokens across 2,365 pages of legal documentation. The questions were designed in a practitioner-oriented, compliance-focused format to reflect realistic operational scenarios. All documents correspond to currently applicable legislation, ensuring contemporary relevance for legal and logistics decision support.

Table 2 presents a sample entry illustrating the dataset structure. Each record includes (i) the question in Polish with an English translation, (ii) a verified answer, (iii) the governing legal instrument (e.g., Regulation (EC) 561/2006, CMR, ADR), (iv) the relevant article or section, (v) page reference, and (vi) an exact supporting quotation serving as the gold-standard citation. The dataset is grounded directly in official EU conventions, regulations, and Polish

transport law, enabling citation-based retrieval, explainable responses, and human-in-the-loop validation. The dataset will be made publicly available upon acceptance.

### 3.2. Retrieval-Augmented Generation System Architecture

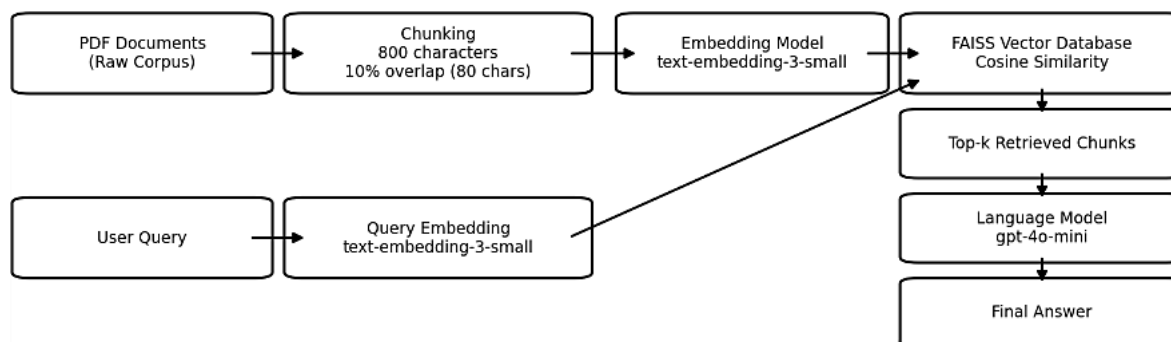
The corpus was processed through a standardized ingestion pipeline within a RAG architecture (Figure 1). Our approach adopts a practical RAG stack designed for assistive use in real operational settings. We employ Faiss (Douze et al., 2025) as the vector database for fast document retrieval, OpenAI's text-embedding-3-small model to generate text embeddings, and the gpt-4o-mini model as the language model, with document ranking based on cosine similarity. Faiss is an open-source library for efficient similarity search over large collections of dense vectors, enabling fast approximate nearest-neighbor retrieval of relevant documents at scale (Douze et al., 2025).

The legal documents were loaded from the project directory in heterogeneous formats (including PDF, DOCX, TXT, and spreadsheet files) and converted into a unified representation as LangChain Document objects. Documents were segmented using a recursive character-based chunking strategy with a maximum chunk length of 800 characters and a 10% overlap (80 characters) to preserve semantic continuity across chunk boundaries. Although Figure 1 illustrates the retrieval pipeline using PDF documents for simplicity, the actual corpus included PDF, DOCX, TXT, and spreadsheet files that were converted into a unified document representation prior to indexing.

Each chunk was assigned a stable identifier derived from document name, page number, and chunk index, enabling deduplication and incremental index updates. This process yielded approximately 8,000 text chunks across the corpus. Each chunk was embedded using OpenAI's text-embedding-3-small model (OpenAI, 2024a), producing 1536-dimensional dense vector representations. Embeddings were indexed in a Faiss vector store (IndexFlatL2) configured for exact nearest-neighbor search. Retrieval was conducted using cosine similarity, with the top-k most similar chunks returned per query ( $k = 5$  in reported experiments).

While the system also supports alternative retrieval configurations, such as similarity score thresholding (threshold = 0.8) and maximum marginal relevance, these were not employed unless explicitly stated. The retrieved passages were supplied to the gpt-4o-mini language model (OpenAI, 2024b) for response generation, ensuring outputs were grounded in semantically relevant source material.

This design emphasizes deterministic retrieval behavior, reproducible indexing across heterogeneous legal sources, and robust semantic matching suitable for real-world regulatory assistance workflows.



**Figure 1.** Proposed RAG-Based Legal Question Answering Architecture.

Source: Own materials.

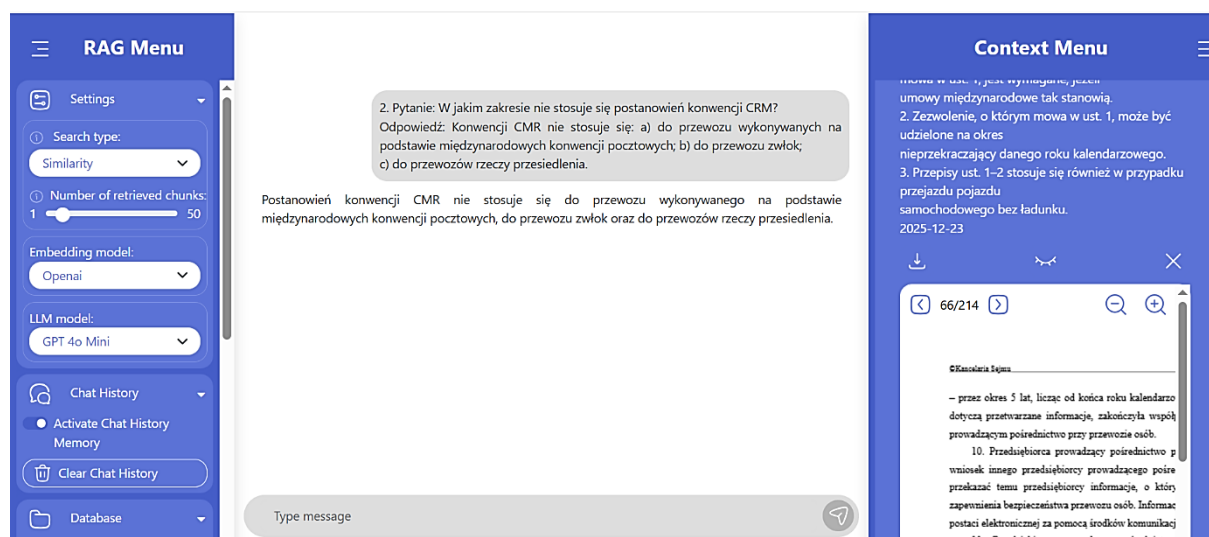
### 3.3. Evaluation Framework

Figure 2 illustrates the user interface of our proposed RAG system in its Polish-language configuration. The interface is designed to support transparent and interpretable legal information retrieval, with particular emphasis on document traceability and evidence-grounded response generation. Rather than functioning as a black-box QA system, the interface enables users to observe how retrieved legal sources contribute to the final response, thereby supporting verification and informed decision-making.

The interface is organized into three main functional areas. The left sidebar, labelled in the RAG Menu, serves as the configuration panel through which users can control both retrieval and generation parameters. Within this panel, practitioners can select the retrieval strategy (e.g., similarity-based search), adjust the number of retrieved document chunks (Top-k), choose the embedding model used for semantic indexing, and specify the language model responsible for answer generation. Additional controls allow management of the document database and optional chat history functionality. These configurable parameters enable users to adjust retrieval of sensitivity and model behavior, improving controllability, and facilitating reproducible experimentation across different query settings.

The central panel presents the generated response produced by the language model. Answers are generated conditionally on the retrieved document fragments and displayed in Polish, reflecting the system's intended use in operational legal contexts. This area functions as the primary interaction space, where users submit queries and receive synthesized legal explanations grounded in retrieved regulatory content.

The right sidebar, referred to as the Context Menu, provides the transparency layer of the system and plays a central role in interpretability. It displays the source document from which supporting evidence was retrieved, including the document title, the exact page location, and a rendered preview of the original legislative PDF. The specific text passage used during generation is highlighted within the document view, allowing users to directly verify the legal basis of the response. By exposing both retrieval context and supporting evidence, the interface enables traceable reasoning and supports trust in AI-assisted legal information retrieval.



**Figure 2.** RAG User Interface for Legal Answer Generation and Document Traceability.

Source: Own materials.

## 4. Results

This section presents the evaluation results of the proposed RAG-based legal question-answering system. Performance is reported for both one-shot and interactive settings and is assessed using answer correctness, document retrieval accuracy, page-level localization, and citation grounding. The results are presented separately for English and Polish questions and are summarized through quantitative performance metrics.

### 4.1. Evaluation Design

System performance was evaluated through a structured manual assessment designed to capture different stages of RAG, including answer correctness, document retrieval, and citation grounding. Each system response was annotated using a three-level category scheme. The category “0” indicates that the generated answer or retrieved legal source was incorrect or irrelevant to the query.

The category “1” denotes a correct result obtained directly from the original query without any modification. The category “2” represents cases in which a correct answer or supporting legal evidence was obtained only after minor prompt refinement, such as clarifying jurisdictional scope or explicitly referencing a regulation. This distinction enables the evaluation to separate fully automatic system performance from results achievable through realistic user interaction, reflecting the intended assistive use of the RAG system.

## 4.2. Performance Analysis

To assess system behavior under both fully automatic and interactive settings, two complementary experiments were conducted. Experiment 1 evaluates one-shot RAG performance, in which queries were submitted without any prompt modification or iterative interaction on the whole dataset. The results are summarized in Table 3. The “Cases” column reports the number of evaluated QA instances for each dataset condition.

The Answer metric measures the proportion of responses categorized as “1”, representing correct answers produced directly from the original query. Retrieval performance is evaluated separately using the “Document” metric, which measures whether the legally relevant source document was retrieved, and the “Page” metric, which assesses whether the relevant evidence was localized at the correct page level. Finally, “Grounded” represents the strictest evaluation criterion, counting only cases in which the answer was correct and supported by both the correct document and page, indicating fully verifiable responses.

**Table 3.**  
*One-shot RAG performance (without prompt refinement)*

| Condition  | Cases | Answer | Document | Page  | Grounded |
|------------|-------|--------|----------|-------|----------|
| Combined   | 99    | 59%    | 58%      | 31.3% | 26.3%    |
| English QA | 45    | 69%    | 78%      | 42.2% | 33.3%    |
| Polish QA  | 44    | 61.4%  | 37%      | 25%   | 25%      |

Source: Own materials.

Experiment 2 investigates interactive use of the system, acknowledging that RAG workflows are inherently iterative and users may refine prompts to improve results. In this experiment, conducted on a subset comprising approximately half of the dataset, users were allowed to modify the original query when the initial response was incorrect or insufficient. The outcomes of this process are presented in Table 4, which distinguishes between answers obtained directly (correct without refinement) and those achieved only after prompt refinement (correct after refinement), alongside remaining incorrect cases.

Prompt refinement followed a procedure in which the evaluator was allowed to clarify the original query without altering its substantive meaning. Refinements typically involved adding explicit references to relevant legal instruments (e.g., “According to Regulation (EC) No 561/2006”), specifying jurisdictional context (e.g., EU regulation or Polish law), or encouraging retrieval from additional language sources when the legal corpus contained multilingual documents. These refinements were intended to mimic realistic user interaction, where practitioners may reformulate queries to provide clearer regulatory context.

The Total values reported in Table 4 are calculated by combining results obtained without prompt refinement (category= “1”) and those achieved after prompt refinement (category= “2”). Thus, the Total metric captures all cases in which a correct answer was ultimately obtained, either directly from the original query or following minimal query reformulation.

This measure provides an estimate of practical system usability in an interactive decision-support setting. The reported percentages are computed by dividing the sum of these correct outcomes (categories “1” & “2”) by the total number of evaluated cases under each condition, including incorrect responses (category= “0”). The Total row therefore reflects overall system performance after interaction, enabling a direct comparison with the one-shot baseline results presented in Table 3.

Together, these experiments allow separate analysis of (i) baseline automatic performance and (ii) performance achievable through realistic human-in-the-loop interaction. This evaluation design provides a comprehensive assessment of generation accuracy, retrieval effectiveness, and evidential grounding, while also demonstrating the practical benefits of iterative query refinement in legal QA scenarios.

**Table 4.**

*Performance with human-in-the-loop interaction (with prompt refinement)*

|            | Condition                  | Answer | Document | Page  | Grounded |
|------------|----------------------------|--------|----------|-------|----------|
| COMBINED   | Incorrect                  | 9      | 19       | 28    | -        |
|            | Correct without refinement | 32     | 26       | 16    | -        |
|            | Correct after refinement   | 8      | 4        | 5     | -        |
|            | Total                      | 82%    | 61.2%    | 43%   | 41%      |
| ENGLISH QA | Incorrect                  | 2      | 3        | 11    | -        |
|            | Correct without refinement | 18     | 19       | 10    | -        |
|            | Correct after refinement   | 5      | 3        | 4     | -        |
|            | Total                      | 92%    | 88%      | 56%   | 52%      |
| POLISH QA  | Incorrect                  | 7      | 16       | 17    | -        |
|            | Correct without refinement | 14     | 7        | 6     | -        |
|            | Correct after refinement   | 3      | 1        | 1     | -        |
|            | Total                      | 71%    | 33.3%    | 29.2% | 29.2%    |

Source: Own materials.

The comparison between Tables 3 and 4 demonstrates that allowing minimal prompt refinement substantially improves overall system effectiveness across all evaluation dimensions. Under one-shot conditions (Table 3), the system achieved moderate answer accuracy (59% for the combined dataset), with comparable document-level retrieval performance (58%). In contrast, page-level localization (31.3%) and fully grounded responses (26.3%) were substantially lower. This pattern suggests that one-shot prompting may not always provide sufficient guidance for the RAG system to identify and incorporate the most relevant supporting passages. As a result, answers may be generated without clearly grounded evidence, indicating that additional clarification or prompt refinement may be required to obtain fully supported responses.

When interactive refinement was introduced (Table 4), performance increased markedly. Combining answers obtained directly with those achieved after prompt refinement raised answer accuracy to 82% in the combined dataset, alongside improvements in document retrieval (61.2%), page identification (43%), and grounded responses (41%). These results indicate that many initial failures were not caused by missing knowledge but by ambiguity in

query formulation, which could often be resolved through minor clarification of jurisdictional scope or regulatory references. The findings therefore highlight iterative interaction as an inherent component of practical RAG deployment rather than a secondary correction mechanism.

Performance differences between language conditions further support this interpretation. English-language queries consistently achieved higher scores, even when underlying legal materials were available only in Polish. This pattern likely reflects stronger English representation in current embedding and language models rather than limitations of the retrieval framework itself. Nevertheless, meaningful performance was maintained across both datasets, suggesting that general-purpose RAG architectures can operate within multilingual regulatory environments typical of European road logistics without domain-specific fine-tuning.

Qualitative analysis revealed that retrieval behavior was highly sensitive to prompt specificity and linguistic cues. For example, when the question was phrased generically as: “What is the break time after 4.5 hours of driving?”, the system initially retrieved a legally valid but contextually different provision (30 minutes) grounded in broader Polish working-time regulations. However, when the question explicitly referenced “According to Regulation (EC) No 561/2006”, the system correctly retrieved Article 7 of the Regulation and generated the expected answer (45 minutes). This observation illustrates how minor variations in wording and jurisdictional cues can substantially influence retrieval behavior. Importantly, such ambiguity reflects real business practice, where legal and operational questions are often formulated imprecisely even by domain experts operating in their native language, reinforcing the importance of prompt design in assistive legal QA systems.

Similarly, adding the phrase “in Polish law” altered document retrieval behavior, prioritizing national legislation over EU regulations. Increasing the number of retrieved documents (k-value) further influenced citation grounding. These findings indicate that general-purpose RAG systems are highly sensitive to linguistic specification and jurisdictional cues. Rather than constituting system errors, such variations reflect the layered structure of transport legislation and the interaction between the EU and national law. From a practical perspective, this suggests that prompt design plays a critical role in achieving accurate and regulation-specific legal QA.

In several instances, the LLM successfully retrieved the correct document and relevant section but nevertheless failed to extract and report the precise answer. For example, in response to the question, *How much can the liability of the ATA Carnet guarantor exceed customs duties and import taxes?* The expected answer is: According to the provisions of the ATA Convention, it cannot exceed 10%.

However, the model generated the following response: “The guaranteeing association shall not be required to pay a sum greater than the total amount of the import duties and taxes, together with any applicable interest”. While the correct source passage, “The liability of the guaranteeing association cannot exceed the amount of customs duties and import taxes by more

than 10%”, was correctly retrieved and cited in the context window, the model failed to report the critical numerical qualifier (10%) in its final answer. This example illustrates a failure at the answer extraction or synthesis stage rather than at retrieval, where the relevant evidence was available but not accurately incorporated into the generated response.

Overall, the results suggest that lightweight, general-purpose RAG systems can provide practically useful assistance for regulatory QA when used interactively. Performance gains observed under prompt refinement emphasize that legal AI systems should be evaluated not only as static models but as collaborative decision-support tools in which iterative clarification plays a central operational role. Future improvements are therefore likely to depend on enhanced multilingual retrieval, improved prompt guidance, and stronger evidence-aware answer generation mechanisms.

## 5. Discussion

Regulation is often understood primarily as a mechanism that restricts behavior; however, it also plays an enabling role by creating structured and predictable conditions that allow complex systems such as international transport to operate effectively. By addressing market failures and establishing shared rules, regulatory frameworks support coordination, safety, and accountability across actors.

Consequently, understanding regulation becomes an operational necessity in logistics practice, where daily decisions depend on interpreting multilayered legal requirements (Baldwin et al., 2011). The empirical findings of this study highlight that this challenge is increasingly not only legal but informational: the rapid growth of available data and the expanding complexity of regulatory frameworks make effective legal understanding progressively more difficult for practitioners. Even experienced professionals struggle to locate and interpret relevant provisions across large and continuously evolving bodies of law, while information obtained from general online sources often varies in reliability. Although large language models enable natural-language access to legal information, their tendency to produce confident but incorrect responses limits their applicability in high-stakes regulatory contexts. These limitations have motivated the adoption of RAG, which grounds AI responses in retrieved legal sources to improve reliability and reduce hallucinations (Li, 2024).

Against this background, the results demonstrate that regulatory QA can be supported through a deliberately lightweight and operationally oriented RAG architecture. Compared with recent legal RAG-based QA systems, which frequently introduce hybrid pipelines, knowledge graphs, or multi-model reasoning mechanisms, the proposed approach prioritizes transparency, reproducibility, and deployability. Hybrid legal QA agents, for example, integrate RAG with model ensembling and human-in-the-loop updating to improve robustness in judicial

environments (Xi et al., 2025), while survey evidence shows that advanced legal RAG systems increasingly rely on iterative retrieval, knowledge graph augmentation, or fine-tuned generators to enhance reasoning performance (Hindi et al., 2025).

In contrast, our framework employs a single-stage retrieval pipeline using dense embeddings and a general-purpose language model. Rather than representing an optimized or fully tailored legal QA architecture, this configuration serves as a lightweight baseline intended to evaluate whether retrieval-grounded generation can already provide meaningful regulatory assistance in an operational setting. The results indicate that even without architectural extensions such as iterative retrieval, model fine-tuning, or knowledge-graph integration, grounded answers can be produced with reasonable reliability. This suggests that practical compliance-oriented support may be achievable using relatively modest system configurations, while more complex architectures remain a promising direction for future improvement, particularly in environments where higher precision or deeper legal reasoning is required. As the present study was not designed as a controlled multilingual retrieval experiment, the observed differences between English and Polish queries may reflect a combination of embedding-model characteristics, corpus language composition, translation effects, and query formulation strategies.

Moreover, the numerical extraction risk Carnet example demonstrates that retrieval-grounded systems may still generate incorrect responses when critical numerical information is omitted during answer synthesis. In compliance-sensitive environments, such omissions may lead to incorrect operational decisions despite successful retrieval of the relevant legal source. Therefore, future work should investigate mechanisms for improving numerical consistency and evidence-aware answer generation.

From a research perspective, this study contributes a domain-specific Polish legal QA dataset, an expert-validated RAG framework, and empirical insights into how AI systems can complement human expertise in regulatory decision support. The results indicate that general-purpose RAG systems can support transport legal QA without domain-specific fine-tuning, achieving meaningful assisted accuracy across multilingual queries. Performance was consistently higher when questions were formulated in English, even when source documents were available in Polish. Rather than representing a system failure, this pattern reflects the stronger English representation within current embedding and language models and reveals how multilingual semantic spaces and jurisdictional cues influence retrieval behavior. In practice, this finding mirrors the layered interaction between EU-level and national transport legislation, where linguistic formulation implicitly guides regulatory interpretation. Importantly, the system maintained consistent performance across languages, demonstrating that general RAG architectures can operate within multilingual regulatory environments typical of European road logistics.

From an operational standpoint, these results are significant because logistics practitioners rarely require fully formal legal reasoning; instead, they need rapid orientation toward applicable regulations supported by traceable sources. The observed variation between Polish and English queries therefore illustrates not merely technical limitations but the structural complexity of regulatory environments and the importance of prompt formulation in compliance-oriented legal QA. Taken together, the findings suggest that practical AI-assisted compliance support may be achievable without costly domain adaptation, while future improvements are likely to arise from multilingual retrieval optimization and human-in-the-loop interaction design. By demonstrating that grounded legal assistance can be achieved using general-purpose models and modest data resources, this study highlights a pathway toward scalable, transparent, and operationally usable AI systems for regulatory decision support.

## 6. Conclusion

This study demonstrates the feasibility of applying RAG system to operational legal QA in road logistics, an area where practitioners must navigate complex and continuously evolving regulatory frameworks. Using a lightweight architecture and a curated practitioner-oriented dataset, the proposed system produced citation-linked answers that supported regulation discovery and interpretation without requiring large-annotated corpora or domain-specific model fine-tuning. The findings suggest that effective regulatory assistance can be achieved through grounded retrieval rather than increasingly complex model adaptation, although precise citation localization remains a key technical challenge.

By introducing a transport-specific legal QA dataset and empirically analyzing multilingual and prompt-sensitivity effects, the study provides initial evidence that AI-assisted regulation search can help bridge the gap between complex legal frameworks and everyday logistics decision-making. These results highlight the potential of transparent, retrieval-based AI systems as practical compliance support tools and establish a foundation for future human-in-the-loop evaluation and workflow-integrated regulatory assistance systems.

This study has several limitations. The evaluation was conducted on a pilot dataset of 50 questions and focused on a specific regulatory domain and jurisdiction. Furthermore, the system was evaluated in an offline setting rather than through real-world user studies. Despite these limitations, the findings provide evidence that retrieval-grounded AI can support regulatory decision-making through transparent, citation-linked responses. The study contributes new knowledge regarding the role of prompt specificity, multilingual retrieval behavior, and human-in-the-loop interaction in legal QA systems. These insights can inform the design of future compliance-support tools and provide a foundation for larger-scale evaluations in logistics and other regulated industries.

## 7. Future directions

Prior literature consistently highlights a tension between regulatory demands and operational realities in road logistics. Legal and transport regulations are formal, mandatory, and directly influence operational performance, yet their volume, dispersion, and technical language make systematic interpretation difficult in fast-paced settings. As a result, practitioners frequently rely on tacit or experience-based knowledge, which may lead to inconsistent compliance and increased operational risk. Although AI applications in logistics have demonstrated benefits for decision support and process optimization, existing legal QA and retrieval systems remain largely detached from real-world logistics practice and are rarely evaluated in operational environments with domain experts or end users. Consequently, there is still limited empirical evidence regarding how such systems perform in everyday compliance decision-making or how they are perceived by logistics professionals.

Future research should therefore move beyond technical benchmarking and investigate AI-assisted regulation-search tools as socio-technical systems embedded in actual workflows. In particular, studies should incorporate human-in-the-loop and field-based evaluations to examine how practitioners interpret, trust, and act upon AI-generated recommendations. Technology acceptance frameworks provide a suitable theoretical foundation for this line of inquiry. The Technology Acceptance Model explains adoption through perceived usefulness and perceived ease of use (Davis, 1989), while the Unified Theory of Acceptance and Use of Technology extend this perspective by incorporating social and organizational influences (Venkatesh et al., 2003).

These models have been widely applied across logistics and supply chain contexts, including the adoption of digital technologies (Li et al., 2025), IoT-based systems (Katias, 2025), blockchain platforms (Sharma, 2023; Wong et al., 2024), and mobile logistics applications (Persada et al., 2023). Empirical findings consistently show that perceived usefulness, ease of use, and facilitating conditions strongly determine system uptake. For example, Castiblanco et al. (2026) demonstrate that TAM effectively explains logistics workers' acceptance of monitoring technologies. However, only a limited number of studies combine AI-driven decision-support tools with structured user-acceptance evaluation, such as the TAM/UTAUT-based assessment reported by Adayilo et al. (2026). Importantly, to date, no studies have applied such acceptance frameworks to RAG systems or examined their adoption in the specific context of road transport regulation and compliance.

Accordingly, future work should extend the present research by conducting longitudinal user studies and expert evaluations of AI-assisted legal QA systems in road logistics, assessing not only technical accuracy but also usability, trust, learning effects, and integration into daily operational practices. Such investigations would provide stronger evidence of practical value and help ensure that regulation-search technologies effectively bridge the gap between complex

legal requirements and real-world logistics decision-making.

Regarding the RAG system, a promising direction for future research is the integration of structured metadata into the RAG pipeline to improve answer precision, particularly for numerically sensitive regulatory queries. During ingestion, chunks could be enriched with explicit metadata fields capturing quantitative elements such as percentages and time limits. Such structured annotations would enable metadata-aware retrieval, filtering, or re-ranking mechanisms that prioritize passages containing explicit quantitative constraints. This approach may substantially reduce the risk that critical numerical qualifiers are overlooked during answer synthesis, thereby improving factual accuracy and reliability in legal and regulatory assistance systems.

## References

1. Abdallah, A., Piryani, B., Jatowt, A. (2023). Exploring the state of the art in legal QA systems. *Journal of Big Data*, 10(1), 127. <https://doi.org/10.1186/s40537-023-00802-8>
2. Adayilo, D. M., Oyefolahan, I. O., Ndunagu, J. N., Otuya, C., Malcalm, E., Twabu, K. (2026). AI-powered tutoring systems for personalised learning feedback in developing secondary education contexts. *Journal of Future Artificial Intelligence and Technologies*, 2(4), 549–564. <https://doi.org/10.62411/faith.3048-3719-290>
3. Adu, J.P., Dorasamy, N., Keelson, S.A. (2023). Road transport infrastructure and supply chain performance in the beverage manufacturing setting: Does road safety compliance matter? *Journal of Law and Sustainable Development*, 11(3), e581. <https://doi.org/10.55908/sdgs.v11i3.581>
4. Ahmad, W., Chi, J., Tian, Y., Chang, K.W. (2020). *PolicyQA: A reading comprehension dataset for privacy policies*. Findings of the Association for Computational Linguistics: EMNLP 2020, 743–749. <https://doi.org/10.18653/v1/2020.findings-emnlp.66>
5. Akarajaradwong, P., Pothavorn, P., Chaksangchaichot, C., Tasawong, P., Nopparatbundit, T., Pratai, K., Nutanong, S. (2025a). *NitiBench: Benchmarking LLM frameworks on Thai legal question answering capabilities*. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 34304-34327. <https://doi.org/10.48550/arXiv.2502.10868>
6. Akarajaradwong, P., Pothavorn, P., Chaksangchaichot, C., Tasawong, P., Nopparatbundit, T., Pratai, K., Nutanong, S. (2025b). *NitiBench: A comprehensive study of LLM framework capabilities for Thai legal question answering*. arXiv. <https://arxiv.org/abs/2502.10868>
7. Akbar, M., Sumardi, J., Razak, A., Noor, S.M. (2024). Harmonization of the law on supply chain integration of the national logistics system in the transformation of port management

- for supporting sustainable development. *Journal of Lifestyle and SDGs Review*, 5(2), e03265. <https://doi.org/10.47172/2965-730X.SDGsReview.v5.n02.pe03265>
8. Baldwin, R., Cave, M., Lodge, M. (2011). *Understanding regulation: Theory, strategy, and practice*. Oxford University Press. <https://doi.org/10.1093/acprof:osobl/9780199576081.001.0001>
  9. Castiblanco Jimenez, I.A., Olivetti, E.C., Mangano, G., Rafele, C., Moos, S., Vezzetti, E., Federica, M. (2026). A technology acceptance model to explore factors of acceptance of emotional monitoring technologies in the logistics industry. *International Journal of Human-Computer Interaction*, 1-30. <https://doi.org/10.1080/10447318.2025.2611349>
  10. Competition on Legal Information Extraction/Entailment (COLIEE 2024) (2024). *11th competition on legal information extraction and entailment (tasks 1-4)*. <https://sites.ualberta.ca/~rabelo/COLIEE2024/>
  11. Davis, F.D. (1989). Technology acceptance model: TAM. In: M.N. Al-Suqri, A.S. Al-Aufi (Eds.), *Information seeking behavior and technology adoption* (pp. 205-219). University of Michigan. <https://www.scirp.org/reference/referencespapers?referenceid=4109007>
  12. de Vos, I.M.A., van den Boogerd, G.L., Fennema, M.D., Correia, A. (2021). *Comparing in context: Improving cosine similarity measures with a metric tensor*. Proceedings of the 18th International Conference on Natural Language Processing (ICON), 128-138. <https://aclanthology.org/2021.icon-main.17/>
  13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171-4186. <https://doi.org/10.18653/v1/N19-1423>
  14. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., Jégou, H. (2025). *The Faiss library*. arXiv. <https://arxiv.org/abs/2401.08281>
  15. Duan, X., Wang, B., Wang, Z., Ma, W., Cui, Y., Wu, D., ..., Liu, Z. (2019). *CJRC: A reliable human-annotated benchmark dataset for Chinese judicial reading comprehension*. China National Conference on Chinese Computational Linguistics, 439-451. [https://doi.org/10.1007/978-3-030-32381-3\\_36](https://doi.org/10.1007/978-3-030-32381-3_36)
  16. Eurostat (n.d.). *Road freight transport statistics*. European Commission. Retrieved from: [https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Road\\_freight\\_transport\\_statistics](https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Road_freight_transport_statistics), January 15, 2026.
  17. Filippova, T.A., Voronina, S.V. (2021). Organizational and legal aspects of transport logistics as a factor of sustainable development. *IOP Conference Series: Earth and Environmental Science*, 670(1), 012048. <https://doi.org/10.1088/1755-1315/670/1/012048>
  18. Garnowski, K. (2017). Funkcja dowodowa listu przewozowego w lądowym transporcie towarów. *Problemy Transportu i Logistyki*, 40(4), 31-40. <https://doi.org/10.18276/ptl.2017.40-03>

19. Glistau, E., Trojahn, S., Coello Machado, N., Behrendt, F. (2024). Logistics 4.0 and the state of knowledge: Interviews with experts. *Polish Journal of Management Studies*, 30(2), 126–145. <https://doi.org/10.17512/pjms.2024.30.2.08>
20. Hendrycks, D., Burns, C., Chen, A., Ball, S. (2021). *CUAD: An expert-annotated NLP dataset for legal contract review*. arXiv. <https://doi.org/10.48550/arXiv.2103.06268>
21. Hindi, M., Mohammed, L., Maaz, O., Alwarafy, A. (2025). Enhancing the precision and interpretability of retrieval-augmented generation (RAG) in legal technology: A survey. *IEEE Access*, 13, 46171-46189. <https://doi.org/10.1109/ACCESS.2025.3550145>
22. Islam, D.M.Z., Meier, J.F., Aditjandra, P.T., Zunder, T.H., Pace, G. (2013). Logistics and supply chain management. *Research in Transportation Economics*, 41(1), 3-16. <https://doi.org/10.1016/j.retrec.2012.10.006>
23. Karp, M., Kubaszewska, A., Król, M., Król, R., Smywiński-Pohl, A., Szymański, M., Wydmański, W. (2025). *LLM-as-a-judge is bad, based on AI attempting the exam qualifying for the member of the Polish National Board of Appeal*. arXiv. <https://doi.org/10.48550/arXiv.2511.04205>
24. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., Kiela, D. (2021). *Retrieval-augmented generation for knowledge-intensive NLP tasks*. arXiv. <https://arxiv.org/abs/2005.11401>
25. Li, R. (2024). *Enhanced large language models-based legal query responses through retrieval augmented generation*. Proceedings of the 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024), 185, 237. [https://doi.org/10.2991/978-94-6463-512-6\\_27](https://doi.org/10.2991/978-94-6463-512-6_27)
26. Li, Y., Ding, N., Zhao, T., Chen, M. (2025). What drives or hinders the adoption of sustainable smart logistics in rural areas?—A mixed-methods analysis. *Sustainability*, 17(14), 6626. <https://doi.org/10.3390/su17146626>
27. Liu, B., Zhu, Z., Ai, Q., Liu, Y., Wu, Y. (2024). *LeDQA: A Chinese legal case document-based question answering dataset*. Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), 5385-5389. <https://doi.org/10.1145/3627673.3679154>
28. Marchewka, W., Cierzniewska, J. (2025). Trendy w zastosowaniach sztucznej inteligencji w logistyce w erze Przemysłu 4.0. *Młodzi Logistycy w Nauce*, 10. <https://doi.org/10.71433/2025.96.3.01>
29. Martineau, K. (2023). *What is retrieval-augmented generation?* IBM Blog. <https://research.ibm.com/blog/retrieval-augmented-generation-RAG>
30. Matak, M., Chudziak, J.A. (2025). *GAIus: Combining GenAI with legal clauses retrieval for knowledge-based assistant*. Proceedings of the 17th International Conference on Agents and Artificial Intelligence, 868-875. <https://doi.org/10.5220/0013191800003890>

31. Muchori, J.G., Mwangi, P.M. (2025). A systematic literature review on application of text embedding techniques in NLP. *International Journal of Scientific Research in Computer Science and Engineering*, 13(6), 45-61. <https://doi.org/10.26438/ijsrcse.v13i6.791>
32. Mutangili, S.K. (2019). Influence of transportation and logistics law on supply chain performance of energy development agencies in Kenya: Case of Kenya Power and Lighting Company Limited. *Journal of Procurement & Supply Chain*, 3(3), 21-38. <https://stratfordjournalpublishers.org/journals/index.php/journal-of-procurement-supply/article/view/375>
33. Nigam, S.K., Mishra, S.K., Mishra, A.K., Shallum, N., Bhattacharya, A. (2023). *Legal question-answering in the Indian context: Efficacy, challenges, and potential of modern AI models*. arXiv. <https://arxiv.org/abs/2309.14735>
34. Odzimek, T. (2023). Law in logistics. In: *Współczesne wyzwania logistyki* (pp. 167-179). Politechnika Częstochowska. Wydawnictwo Politechniki Częstochowskiej. ISBN: 9788371939266. <https://bibliotekanauki.pl/books/25122832>
35. OpenAI (2024a). text-embedding-3-small (Version 3) [Large language model]. Retrieved from: <https://platform.openai.com/>, February 17, 2026.
36. OpenAI. (2024b). gpt-4o-mini [Large language model]. OpenAI Platform. Retrieved from: <https://platform.openai.com/>, February 17, 2026.
37. Orłowska, M., Chackiewicz, M. (2024). Logistics and customs handling—New technologies and operational efficiency and compliance with international regulations. *Scientific Papers of Silesian University of Technology. Organization & Management [Zeszyty Naukowe Politechniki Śląskiej. Seria Organizacji i Zarządzanie]*, 211. <https://doi.org/10.29119/1641-3466.2024.211.30>
38. Pennington, J., Socher, R., Manning, C.D. (2014). *GloVe: Global vectors for word representation*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532-1543. <https://doi.org/10.3115/v1/D14-1162>
39. Persada, S.F., Afandi, F., Redi, A.A.N.P., Nadlifatin, R., Prasetyo, Y.T., Kurniawan, A.C. (2023). Mix method analysis for analyzing user behavior on logistic company mobile pocket software. *Jurnal Sistem dan Manajemen Industri*, 7(1), 69-81. <https://doi.org/10.30656/jsmi.v7i1.5937>
40. Pipitone, N., Alami, G.H. (2024). *LegalBench-RAG: A benchmark for retrieval-augmented generation in the legal domain*. arXiv. <https://doi.org/10.48550/arXiv.2408.10343>
41. Poliak, M., Mrnikova, M., Simurkova, P., Medvid, P., Poliakova, A., Hernandez, S. (2018). *Social law in road transport as a tool for safety in road transport*. Proceedings of the 2018 XI International Science-Technical Conference Automotive Safety, 1-7. <https://doi.org/10.1109/AUTOSAFE.2018.8373317>
42. Puspadam Katias (2025). Adoption of Internet of Things (IoT) in Supply Chain Management: A Systematic Literature Review. *International Conference on Digital*

- Business Innovation and Technology Management (ICONBIT)*, 1(2). Retrieved from: <https://proceeding.unesa.ac.id/index.php/iconbit/article/view/5934>
43. Quevedo, E., Cerny, T., Rodriguez, A., Rivas, P., Yero, J., Sooksatra, K., ..., Taibi, D. (2023). Legal natural language processing from 2015 to 2022: A comprehensive systematic mapping study of advances and applications. *IEEE Access*, 12, 145286-145317. <https://doi.org/10.1109/ACCESS.2023.3333946>
  44. Rawling, M. J., Shepherd, N., Johnstone, R., Nossar, I. (2017). *Australian supply chain regulation project: Draft report on road transport sector regulation*. University of Technology Sydney. [https://www.uts.edu.au/globalassets/sites/default/files/2017-12/supply-chain-regulation-in-road-transport\\_draft-report.pdf](https://www.uts.edu.au/globalassets/sites/default/files/2017-12/supply-chain-regulation-in-road-transport_draft-report.pdf)
  45. Redelaar, F., Van Drie, R., Verberne, S., De Boer, M. (2024). *Attributed question answering for preconditions in the Dutch law*. Proceedings of the Natural Legal Language Processing Workshop 2024, 154-165. <https://doi.org/10.18653/v1/2024.nllp-1.12>
  46. Rybak, P., Przybyła, P., Ogrodniczuk, M. (2024). *PolQA: Polish question answering dataset*. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 12846-12855. <https://aclanthology.org/2024.lrec-main.1125/>
  47. Sharma, A., Sharma, A., Singh, R.K., Bhatia, T. (2023). Blockchain adoption in agri-food supply chain management: An empirical study of the main drivers using extended UTAUT. *Business Process Management Journal*, 29(3), 737-756. <https://doi.org/10.1108/BPMJ-10-2022-0543>
  48. Siino, M., Falco, M., Croce, D., Rosso, P. (2025). Exploring LLMs applications in law: A literature review on current legal NLP approaches. *IEEE Access*. <https://doi.org/10.1109/ACCESS.2025.3533217>
  49. Smywiński-Pohl, A., Król, M., Mirosław, P. (2025). *From statement of facts to statutory provisions—Efficient retrieval of relevant legislation*. International Conference on Computational Science, 258-272. <https://www.iccs-meeting.org/archive/iccs2025/papers/159110250.pdf>
  50. Ujwal, U., Surampudi, S.S.H., Mitra, S., Saha, T. (2024). “Reasoning before responding”: Towards legal long-form question answering with interpretability. Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, 4922-4930. <http://hdl.handle.net/10150/679213>
  51. Venkatesh, V., Morris, M.G., Davis, G.B., Davis, F.D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3), 425-478. <https://doi.org/10.2307/30036540>
  52. Walther, J. (2008). Transport law and forwarder law. In: *Key aspects of German business law: A practical manual* (pp. 217-237). Springer. <https://doi.org/10.1007/978-3-540-68577-7>

53. Wang, S., Scardigli, A., Tang, L., Chen, W., Levkin, D., Chen, A., ..., Hendrycks, D. (2023). *MAUD: An expert-annotated legal NLP dataset for merger agreement understanding*. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 16369-16382. <https://doi.org/10.18653/v1/2023.emnlp-main.1019>
54. Wiratunga, N., Abeyratne, R., Jayawardena, L., Martin, K., Massie, S., Nkisi-Orji, I., Weerasinghe, R., Liret, A., Fleisch, B. (2024). CBR-RAG: Case-based reasoning for retrieval augmented generation in LLMs for legal question answering. In: J.A. Recio-Garcia, M.G. Orozco-del-Castillo, D. Bridge (Eds.), *Case-based reasoning research and development. Lecture Notes in Computer Science, Vol. 14775* (pp. 445-460). Springer. [https://doi.org/10.1007/978-3-031-63646-2\\_29](https://doi.org/10.1007/978-3-031-63646-2_29)
55. Wojtasik, K., Domagała, A., Oleksy, M., Piasecki, M. (2025). Towards complex question answering in Polish language. In: *Computational collective intelligence*. 17th International Conference, ICCCI 2025, Ho Chi Minh City, Vietnam, November 12-15, 2025, proceedings, part I, pp. 256-268. Springer. [https://doi.org/10.1007/978-3-032-09318-9\\_18](https://doi.org/10.1007/978-3-032-09318-9_18)
56. Wong, S., Yeung, J.K.W., Lau, Y.Y., Kawasaki, T., Kwong, R. (2024). A critical literature review on blockchain technology adoption in supply chains. *Sustainability*, 16(12), 5174. <https://doi.org/10.3390/su16125174>
57. Xi, Y., Bai, Y., Luo, H., Wen, W., Liu, H., Li, H. (2025). *Hybrid retrieval-augmented generation agent for trustworthy legal question answering in judicial forensics*. arXiv. <https://doi.org/10.48550/arXiv.2511.01668>
58. Yao, R., Wu, Y., Wang, C., Xiong, J., Wang, F., Liu, X. (2025). *Elevating legal LLM responses: Harnessing trainable logical structures and semantic knowledge with legal reasoning*. arXiv. <https://doi.org/10.48550/arXiv.2502.07912>
59. Zhai, H. (2025). *Law GraphRAG: An advanced legal question-answering system*. Proceedings of the 2025 5th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA), 1407-1410. <https://doi.org/10.1109/AIITA65135.2025.11047851>
60. Zheng, L., Guha, N., Anderson, B.R., Henderson, P., Ho, D.E. (2021). *When does pretraining help? Assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings*. Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, 159-168. <https://doi.org/10.1145/3462757.3466088>
61. Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., Sun, M. (2020). *JEC-QA: A legal-domain question answering dataset*. Proceedings of the AAAI Conference on Artificial Intelligence, 34(5), 9701-9708. <https://doi.org/10.1609/aaai.v34i05.6519>