**UDC 004.891.2**

# INTEGRATION OF KNOWLEDGE GRAPHS AND RAG FOR LEGAL NORMS ANALYSIS IN KAZAKHSTAN

**Alen Tassymov**
master's student, Astana IT University, Astana, Kazakhstan
Scientific supervisor: **Bolatzhan Kumalakov**
Ph.D., Associate Professor, Astana IT University, Astana, Kazakhstan

## Introduction

The rapid expansion of digital government services and the increasing complexity of legal regulation have created an urgent demand for intelligent, accurate, and user-friendly tools for semantic search and analysis of normative legal acts.

Recently, large language models (LLMs) have been increasingly applied in the legal field, demonstrating high effectiveness in text generation, question answering, and document analysis. But processing legal acts using LLM training is complicated by the specifics of legal terminology, large volumes of data, high training costs, difficulties in interpreting highly specialized terms, the need for constant updates due to frequent changes in legislation, especially in language groups that initially lack a large amount of data or models trained on them [1]. LLMs often lack specialized knowledge required for accurately solving legal tasks, and their outputs may contain errors or "hallucinations," making careful verification of the information they provide essential [2].

The integration of large language models with vector embeddings, legal knowledge graphs, and Retrieval-Augmented Generation (RAG) introduces a new approach to addressing legal tasks. Embedding models enable the retrieval of semantically relevant fragments of laws even for differently phrased queries, while knowledge graphs preserve the structure of documents and the relationships between norms, allowing the context necessary for LLMs to provide relevant answers. This allows the model to access up-to-date, accurate data in real-time and provide source references, rather than relying solely on its training data [3]. This approach improves search accuracy, reduces the risk of hallucinations, and ensures information relevance without the need for constant retraining of LLMs when laws change [4].

The relevance of this research is determined by the need to improve the efficiency of access to the regulatory legal framework of the Republic of Kazakhstan and to optimize the processes of searching for and interpreting legal norms. With the increasing volume of regulatory legal acts, the complexity of their analysis grows, which requires the application of modern information processing methods and intelligent decision-support systems.

This study compares a retrieval-augmented approach for working with regulatory legal acts, based on data from the Reference Control Bank of Regulatory Legal Acts, with a method based on training large language models on a corpus of court decisions. The proposed approach integrates vector search, structured knowledge representation through knowledge graphs, and generative models to enable semantic retrieval and interpretation of legal information while preserving the relationships and hierarchy of legal norms.

Thus, the research aligns with modern trends in the development of artificial intelligence technologies in the legal domain and aims to improve the accessibility, accuracy, and efficiency of user interaction with regulatory legal information.

## Methods

In this study, we assess whether a retrieval-augmented approach combining vector search with structured knowledge representation through knowledge graphs can match the quality of a specialized, fully trained neural network. Specifically, we aim to evaluate whether an untrained model, when provided with up-to-date legal information via RAG and knowledge graphs, can

generate accurate and legally sound responses comparable to those produced by a dedicated LLM trained on legal corpora.

For quality comparison, each response was rated on a five-point scale according to three criteria: accuracy, completeness, and validity. The evaluation scale is presented in Figure 1.

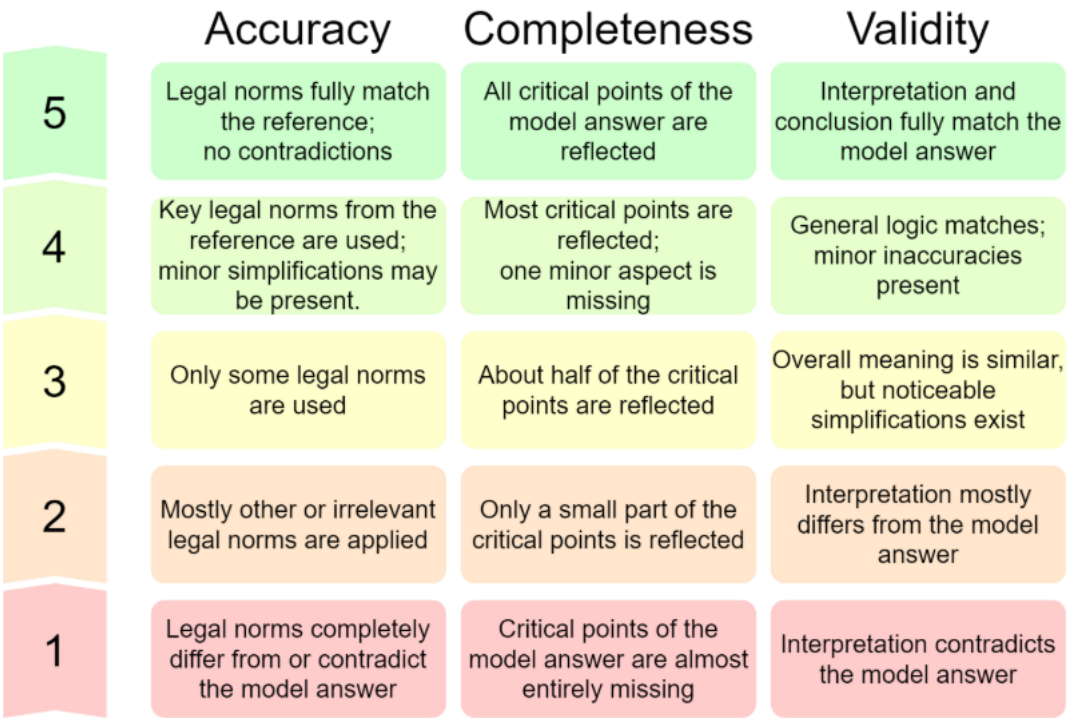


Figure 1. Criteria for assessing accuracy, completeness, and validity.

To compare the quality of the responses, each response received an overall score computed as the sum of the scores for accuracy, completeness, and validity, with a maximum possible score of 15. The evaluation included the average score for each criterion, the average total score per question, and the number of responses with total scores in the ranges of 1–5, 6–10, and 11–15.

The evaluation was performed using the LLM-as-a-judge approach with the Gemini 3.1 Pro model [5]. The model was supplied with detailed instructions and evaluation criteria for assessing the answers. The temperature parameter was set to 0.1 to promote consistent outputs, and high-reasoning mode was enabled.

The evaluation dataset consisted of 20 complex open-ended legal questions covering multiple areas of law. The questions and their reference answers were collected from an online legal question-and-answer service. On this platform, responses are prepared by law students under the supervision of experienced professors. Each reference answer provides a detailed explanation, including citations to relevant legal sources and comprehensive legal reasoning.

To assess the relevance of the legal norms applied, a set of 10 questions was developed, covering various domains of legally regulated relations. These questions require, for their resolution, the application of legal norms enacted as of the beginning of 2026 and not previously utilized in legal practice. The questions were constructed based on the content of the relevant legal provisions and necessitate their explicit reference in the responses. The evaluation of results was conducted based on the presence of these norms in the answers.

As a baseline for evaluating the capabilities of the RAG approach with a knowledge graph, the legal consultation service LegalExpert.kz was selected. This service represents one of the most well-known publicly available solutions for legal analysis in Kazakhstan. The neural network used in this system is trained on more than 100,000 annotated court decisions and is designed to provide consultations on legal matters.

In recent years, there has been increasing interest in hybrid approaches that combine knowledge graphs with RAG, which are actively studied and applied across various domains

[6][7]. However, in the context of Kazakhstan's legal system, the predominant approach still relies on training neural networks on large volumes of legal data. Therefore, LegalExpert.kz serves as a representative example of this traditional approach and provides a suitable baseline for comparison.

Neo4j was selected as the graph database for constructing the legal knowledge graph. Relationships were decomposed into subject-predicate-object triples and represented in the graph as nodes, edges, and their associated properties [8]. This modeling approach provides high flexibility, as nodes can store a variety of properties depending on the structure of the legal knowledge being represented. Neo4j's Cypher query language enables expressive graph queries, facilitating efficient retrieval and extraction of enriched structured information [9].

The hierarchical structure of normative legal acts and their clear organization into documents and sections allow legal norms to be represented as a knowledge graph, helping address common challenges in knowledge graph modeling [10]. Relationships between nodes can model both document structure and broader connections between legal entities [11], while directed edges enable tracing paths from individual paragraphs to relate entities (Figure 2).

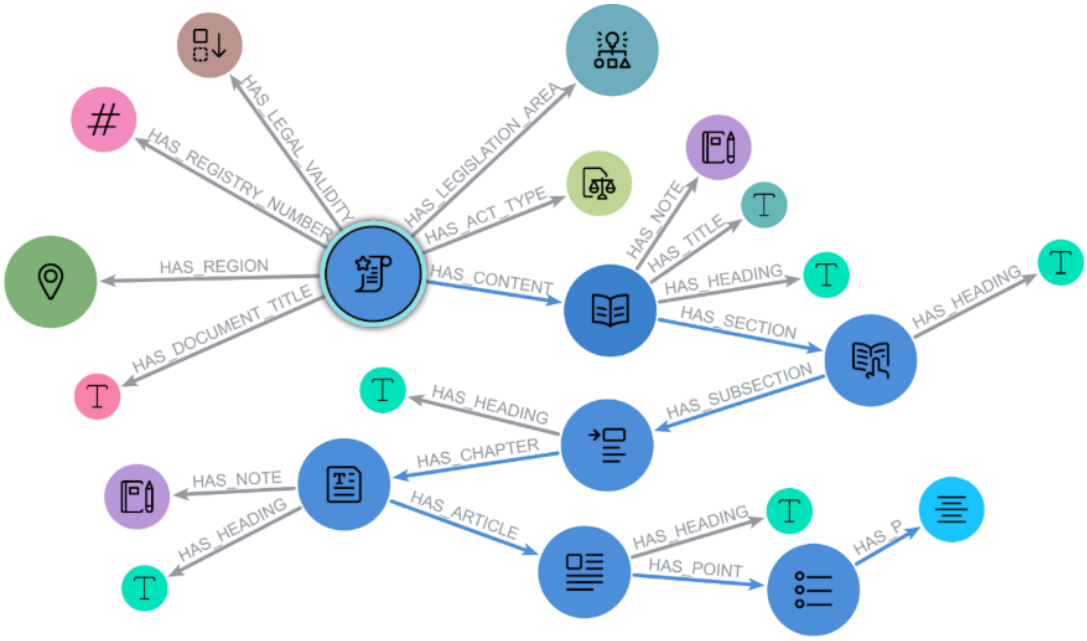


Figure 2. Structure and relationships of the Civil Code of the Republic of Kazakhstan.

The source data for the formation of the knowledge graph was taken from the Reference Control Bank of Normative Legal Acts of the Republic of Kazakhstan in electronic form. The data is presented in JSON format. It contains metadata for identifying the normative legal act, dates of interactions with the act, government agencies related to the act, document status, publication sources, legal validity, region of application, areas of legislation, references to amendments, document language, date of first publication, current version, number of versions, version date, and the content of the normative legal act for its representation in a human-readable form.

For further work with the data, it was decided to first convert it into XML format, as it is widely used as a format for exchanging legal information and is more convenient for the subsequent creation of the knowledge graph [12]. For the study, 309 legal acts were retrieved from the law.gov.kz service and exported in XML format, preserving the metadata of the document hierarchy. The terminal nodes of this structure are text nodes containing the minimal semantic value within the context of the document. These nodes are vectorized into normalized dense embeddings of size 768 using the embeddinggemma-308m model. On an Nvidia 4060 8GB GPU and 16GB of RAM, the vectorization process took 18 minutes with a batch size of 16. In total, 194,539 text nodes were vectorized. These nodes are stored in a local vector database (ChromaDB) along with metadata such as node ID, text representation and source legal document.

Neo4j graph database is deployed on the local machine, where all the documents are sequentially exported, and one-way relationships are created in parallel. This approach facilitates the future aggregation of data to create context for the generator.

Due to the complex structure of legal norms in Kazakhstan, effective data segmentation is essential for RAG systems [13]. In this approach, paragraphs – the smallest structural units of legal acts – are used as encoded text, while a knowledge graph is queried after vector retrieval to enrich the context with a set of relevant articles (Figure 3).

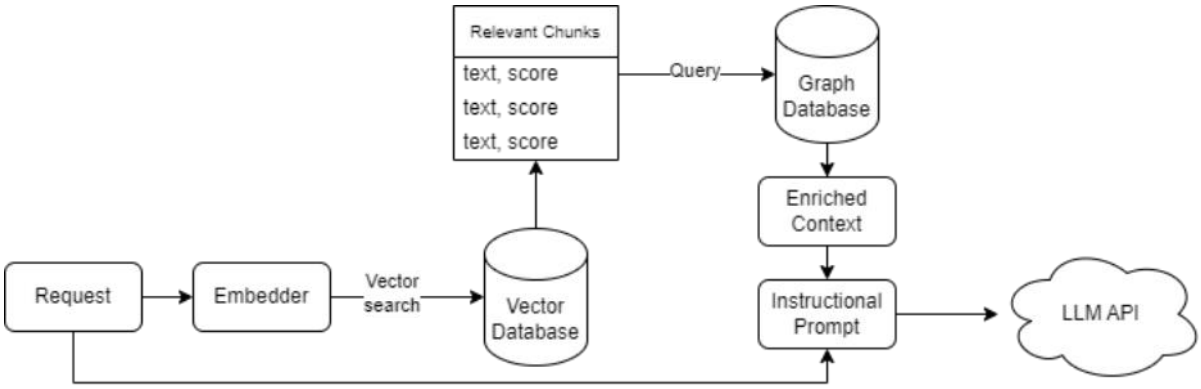


Figure 3. Enriching legal text context using knowledge graphs.

The incoming query in natural language is encoded into a vector representation using embeddinggemma-308M into a 768-dimensional vector. By calculating the cosine similarity, the top 150 closest vectors are retrieved. This size was selected experimentally and is determined by the number of similar text node entries that may contain relevant information within this range. The retrieved vectors are further reranked using the Qwen3 Reranker 0.6B cross-encoder, after which the top 10 vectors are selected for each legal document. The vector IDs are aligned with the data IDs in Neo4j and can be mapped in a Cypher query.

The result of the aggregation in Neo4j consists of a set of complete structural units of the legal act, supplemented with the document's metadata. This query is then combined into a prompt with instructions for the generator (Figure 4).

Role | Rules | Output format | Context | Querry

Prompt

```
You are a legal assistant working based on the provided context of the legislation of the Republic of Kazakhstan.
Rules:
1. Base your answer only on the provided context.
2. Try to provide a useful and practical answer to the user.
3. Do not invent legal norms, articles, or clauses.
4. If the information in the context is incomplete: use only confirmed provisions, logically connect them, but do not add new legal norms.
5. If there is no direct answer, explain which provisions from the context may apply to the situation.
6. Be sure to include references to legal provisions (article, clause, document) if they are present in the context.
7. Do not use knowledge outside the provided context.
8. Do not write the phrase “insufficient information” if it is possible to provide a partial or general conclusion based on the available norms.
9. Carefully analyze the question and identify the sequence of events and the specific subjects of the legal relationship.
Answer structure:
- brief explanation of the situation
- applicable legal provisions (with references) from the context
- practical conclusion with references to legal provisions from the context
- direct answer to the user’s question based on the practical conclusion
User question:
{user_query}
Context:
{context}
```

Figure 4. Generating prompt structure with example.

The Gemini 3.1 Flash Lite was selected as the generator. Queries are sent through the Gemini API. The number of output tokens for the generator is set to 4096, with a temperature of 0.2 and top-p of 1.0.

The generated response is evaluated by an LLM judge against the reference answer using the defined evaluation criteria. The same evaluation procedure is applied to the response obtained from LegalExpert.kz to enable further comparison of the results.

**Results**

After comparison with the reference answers using the LLM judge, the quality of the responses generated by the models was analyzed across three criteria: accuracy, completeness, and validity (Figure 5). The Knowledge Graph–based RAG approach achieved higher scores in 5 out of 20 questions for accuracy and in 6 out of 20 questions for both completeness and validity. When considering all three criteria simultaneously, the Knowledge Graph RAG system produced better overall responses in 3 out of 20 questions.

The LegalExpert neural network achieved higher scores in 4 out of 20 questions for each criterion (accuracy, completeness, and validity). In terms of the combined evaluation across all three metrics, it also produced better responses in 3 out of 20 questions.

Identical scores were observed in 11 questions for accuracy and in 10 questions for both completeness and validity.

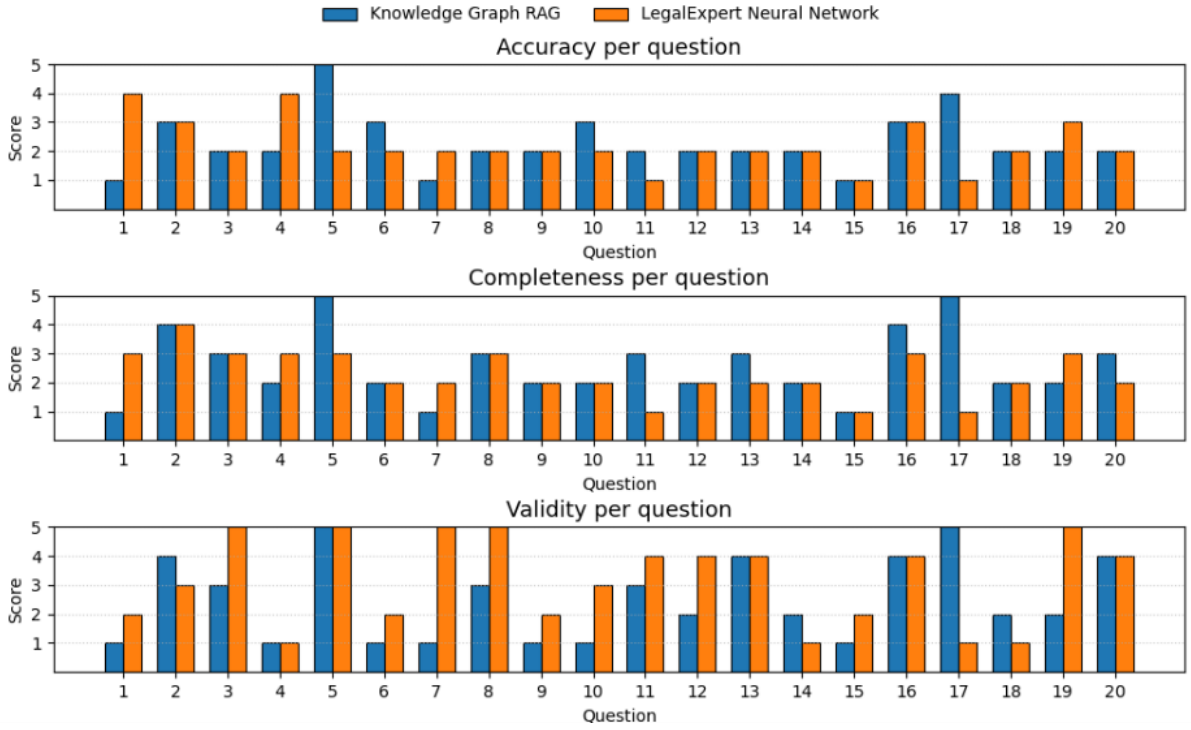


Figure 5. Response quality comparison per question.

The Knowledge Graph RAG system achieved an average total score of 7.4, compared to 6.75 for the LegalExpert neural network when summing the scores across the three evaluation criteria. This result is also reflected in the distribution of total scores: the proportion of responses in the 1–5 range was 5% lower, while the proportion in the 11–15 range was 10% higher for the Knowledge Graph RAG system (Figure 6).

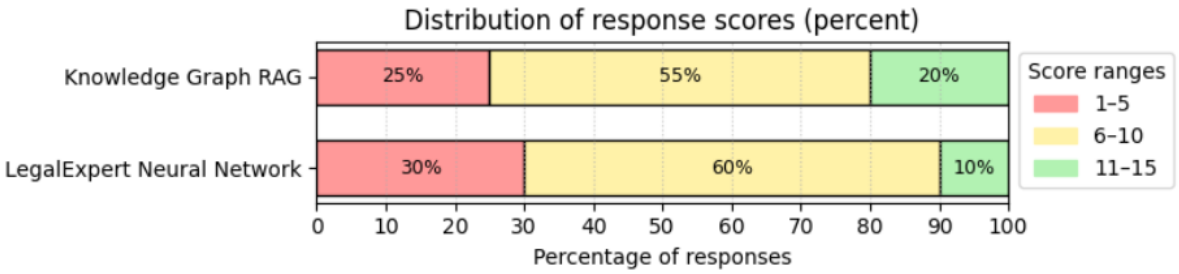


Figure 6. Distribution of response scores in percent.

Figure 7 shows the average scores across the three evaluation criteria. The Knowledge Graph RAG system demonstrates higher average values for accuracy, completeness, and validity compared to the LegalExpert neural network. The differences amount to 4.55% for accuracy, 13.05% for completeness, and 11.11% for validity.

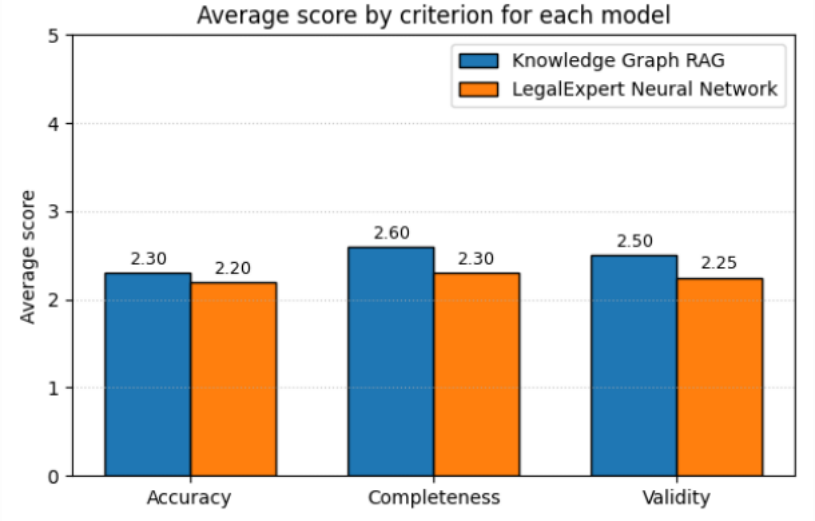


Figure 7. Average score by accuracy, completeness, and validity.

According to the results of the comparison of the relevance of the legal norms applied, the Knowledge Graph RAG system provided correct answers to all 10 out of 10 questions, relying on newly enacted legal norms. In contrast, the LegalExpert neural network failed to address any of the questions, citing the absence of applicable legal provisions.

The results indicate that the Knowledge Graph RAG system and the LegalExpert neural network perform at a comparable level in terms of accuracy, completeness, and validity. However, the Graph RAG system demonstrated a substantial advantage in the relevance of the legal norms applied, while also exhibiting slight advantages in average overall scores and a higher proportion of high-scoring responses, particularly for completeness and validity. These observations indicate that the integration of vector-based retrieval with structured knowledge representation through knowledge graphs enables the generation of responses that are competitive with those of a specialized neural network, while providing a significant advantage in terms of the ability to rapidly integrate newly introduced legal provisions. Moreover, the structured representation of legal norms appears to facilitate a more effective capture of both semantic and structural characteristics of legal texts. Overall, the findings support the potential of hybrid RAG approaches as a viable alternative to fully trained, task-specific neural networks in the domain of legal question answering.

**Conclusion**

This study compared the quality of a retrieval-augmented approach combining vector search and structured knowledge representation through legal knowledge graphs with a specialized

neural network trained on a large corpus of court decisions. The evaluation, performed using an LLM-as-a-judge framework, assessed response quality across three criteria: accuracy, completeness, and validity. Notably, the Knowledge Graph RAG system also exhibited a substantial advantage in terms of the relevance of the legal norms applied, successfully incorporating newly enacted provisions that had not yet been used in practice. Results show that both approaches achieve broadly comparable quality, with the Knowledge Graph RAG system demonstrating slight advantages in average total scores and a higher proportion of high-scoring responses, while also offering the significant benefit of enabling timely updates to the underlying legal knowledge base.

The findings indicate that integrating vector retrieval with structured legal knowledge can provide responses competitive with a dedicated, fully trained neural network, while also leveraging the inherent structure and relationships of legal norms. This approach allows up-to-date and contextually accurate information retrieval without the need for extensive model retraining. Overall, the study highlights the potential of hybrid RAG-based systems as a practical tool for legal question answering, offering an alternative solution to traditional neural network approaches in the context of the Kazakhstan legal system.

**List of sources used**


1. Marco Siino, Exploring the use of LLMs in the Italian legal domain: A survey on recent applications, Computer Law & Security Review, Volume 58, 2025, 106164, ISSN 2212-473X
2. Qian Liu, Hang Yu, Qiqi Wang, Qi Xu, Jinpeng Li, Zhuoqun Zou, Rui Mao, Erik Cambria, Legal knowledge infusion for large language models: A survey, Information Fusion, Volume 125, 2026, 103426, ISSN 1566-2535
3. Amir Abbas Kamalipour, Shahrokh Asadi, Mohammad Mahyar Amiri Chimeh, From vectors to knowledge graphs: A comprehensive analysis of modern retrieval-augmented generation architectures, Computer Science Review, Volume 61, 2026, 100925, ISSN 1574-0137
4. George Hannah, Rita T. Sousa, Ioannis Dasoulas, Claudia d'Amato, On the legal implications of Large Language Model answers: A prompt engineering approach and a view beyond by exploiting Knowledge Graphs, Journal of Web Semantics, Volume 84, 2025, 100843, ISSN 1570-8268
5. Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, Jian Guo, A survey on LLM-as-a-Judge, The Innovation, 2026, 101253, ISSN 2666-6758
6. Shucheng Huang, Chen Sun, Minghao Ning, Yufeng Yang, Changye Ma, Jiaming Zhong, Keqi Shu, Freda Shi, Amir Khajepour, DriveLegal: Toward legally compliant driving via trustworthy hybrid retrieval-augmented LLMs, Expert Systems with Applications, Volume 314, 2026, 131593, ISSN 0957-4174
7. Taoufiq El Moussaoui, Chakir Loqman, From text to knowledge graph: Enhancing arabic legal research with FaCitaGraphAR and AraBERT-driven citation recognition, Knowledge-Based Systems, Volume 335, 2026, 115098, ISSN 0950-7051
8. Jiajie Zhou, Xinyi Liu, Tao Wang, A Study on the Construction of Extenics Database Based on LLM+Neo4j - An Example of Plant Landing Designs, Procedia Computer Science, Volume 266, 2025, Pages 252-260, ISSN 1877-0509
9. Florian Holzschuher, René Peinl, Querying a graph database – language selection and performance considerations, Journal of Computer and System Sciences, Volume 82, Issue 1, Part A, 2016, Pages 45-68, ISSN 0022-0000
10. Valentina Presutti, Enrico Motta, Marta Sabou, Opportunities for Knowledge Graphs in the AI landscape — An application-centric perspective, Journal of Web Semantics, 2025, 100867, ISSN 1570-8268
11. George Hannah, Rita T. Sousa, Ioannis Dasoulas, Claudia d'Amato, On the legal implications of Large Language Model answers: A prompt engineering approach and a view

beyond by exploiting Knowledge Graphs, Journal of Web Semantics, Volume 84, 2025, 100843, ISSN 1570-8268

12. Andrea Colombo, Anna Bernasconi, Stefano Ceri, An LLM-assisted ETL pipeline to build a high-quality knowledge graph of the Italian legislation, Information Processing & Management, Volume 62, Issue 4, 2025, 104082, ISSN 0306-4573

13. Matteo Buffa, Alfio Ferrara, Sergio Picascia, Davide Riva, Silvana Castano, Enhancing legal document building with Retrieval-Augmented Generation, Computer Law & Security Review, Volume 59, 2025, 106229, ISSN 2212-473X