

Review Article

An Explainable Rule-Based Framework for Knowledge Extraction and Legal Knowledge Graph Construction from Arabic Higher Education Legislation

Ghassan Khazal Ali

College of Engineering, University of Diyala, IRAQ.



Article Information

Article history:

Received: 29, 02, 2026

Revised: 22, 06, 2026

Accepted: 14, 07, 2026

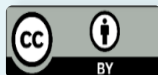
Published: 30, 09, 2026

Keywords:

Knowledge graph,
Information retrieval,
Arabic Legal NLP,
Knowledge Extraction.

Abstract

The volume and complexity of legislative documents are increasing and have significant challenges for legal information management, retrieval, and intelligent decision support. Thus, the technologies related to natural language processing (NLP) and knowledge graph (KG) were used to improve the knowledge extraction in scientific and general domain text but are limited in Arabic legal documents. This limitation is evident in higher education legislation, where these documents have hierarchical structure, frequent changes, extensive cross reference, and special terminology. All of these factors make automated interpretation and knowledge extraction complicated. This work presents an automated framework for constructing an interpretable legal knowledge schema from the Iraqi Ministry of Higher Education legislation, using transparent information extraction based on Python. This framework was developed using a legislative dataset comprising 1930 regulations, laws, ministerial directives, and administrative decisions issued by the ministry of higher education between 1970 and 2025. The proposed work integrates legislative meta data with the full text of the article through many stages of pipelining like documents preprocessing, segmentation into article and clause, legal entity identification, reclassification extraction, and automatic graph construction. The resultant graph provides a structured representation of legislative knowledge that supports legal search, analysis, exploration, and serves as a foundational resource for future graph retrieval-augmented generation (GraphRAG) systems and legal question-answering applications. We conducted experiments on the complete legislative corpus, producing a legal knowledge graph comprising 10,312 article units, 16,597 clauses, 510 normalized entity nodes, and 12,289 extracted relation records. To evaluate the extraction quality, a manually annotated benchmark derived from the first 100 legislative documents was constructed. The proposed framework achieved promising retrieval performance, with entity extraction obtained (precision = 0.901, recall = 0.868, F1 = 0.884), while relation extraction obtained (precision = 0.736, recall = 0.582, F1 = 0.650).



Copyright: © 2026 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License. With the license CC-BY, authors retain the copyright, allowing anyone to download, reuse, re-print, modify, distribute, and/or copy their contribution. The work must be properly attributed to its author.

Corresponding Author: Ghassan Khazal Ali, Department of Computer Engineering, College of Engineering, University of Diyala, Baqubah, Diyala Iraq.



1. INTRODUCTION

Legal norms and legislative acts in the Ministry of Higher Education possess a sophisticated structure comprising laws, regulations, instructions, Council of Ministers decisions, judicial decisions. These documents feature many cross references, temporal provisions, exceptions, and institutional responsibilities. This makes legal search, compliance checking, and advisory activities in the setting of operational efficiency a tedious job for lawyers, higher education institutions, administrators, and students alike. A legal knowledge graph framework represents a structured alternative to raw document collections. Using documents as nodes and their relationships, one can express not only the presence of a certain legislative act but also its structure, cross-references to other legal acts, temporal validity, and legal normativity.

This is especially crucial in the sphere of higher education, where the answer to a certain question might depend on an article, version, or amendment trail rather than the whole document. In this study, we attempt to tackle the problem using a reproducible and audit-ready rule-based pipeline working with the existing corpus and a defined evaluation benchmark. Contrary to the presentation of Python as a scientific novelty, the current study focuses on the traceability of legal segmentation, domain-related rules for extracting legal information, preserved-provenance relations records, and knowledge graphs ready for further inspection and legal retrieval.

The current implementation is rule/symbolic-based; neural language models and GraphRAG are considered as potential additions rather than evaluated systems. Scientific novelty in this paper lies in the combination of the following four rarely combined elements for the extraction of Arabic higher-education legislation: (i) deterministic parsing of articles and clauses, (ii) domain-related extraction of legal references, institutions, authorities, and



normative relations, (iii) provenance-preserved graph construction where relations keep track of document and passage identifiers, and (iv) exact match evaluation protocol distributed alongside with the project.

2. RELATED WORK

Prior work has shown that legal and regulatory texts can be converted into graph-based representations. The work on automated creation of a legal knowledge graph for violence-against-women legislation demonstrated the feasibility of transforming legislation into structured knowledge resources [1]. More general analyses of legal documents using knowledge graphs also argue that graph structure improves traceability and downstream analytics [2]. A second approach to research focuses on automated extraction. Complex knowledge mining pipelines and transfer-learning approaches show that legal entities and relations can be mined from unstructured documents with reasonable accuracy [3]. Knowledge-enhanced large language models extend this direction by combining language-model flexibility with domain-specific legal signals [4]. A third approach focuses on reasoning and compliance in “ontology-driven GraphRAG,” “policy-context alignment,” “OWL-based compliance,” and “deterministic legal agents”; these emphasize that regulatory intelligence must go beyond retrieval and support auditable, temporal, and rule-sensitive reasoning [5, 6, 7, 8]. A fourth one is particularly relevant to this project: retrieval systems for academic regulations. REBot and CatRAG show that graph-enriched retrieval substantially improves advisory interactions over higher-education rules [8]. Complementary work on retrieval-augmented legal intelligence and legal reasoning paths shows that graph structure can improve explainability and legal grounding [9, 10]. Finally, formalization work in deontic logic reminds us that obligations, permissions, and prohibitions require careful modeling if a system is expected to support compliance-sensitive use cases [11].

3. DATASET AND PROBLEM SETTING

The current corpus is based on 1,930 (Table 1) legislative documents. In the JSONL document, there is a unique doc_id, original Arabic text, normalized text, and source field if available. It consists of Iraqi legislative documents related to higher education and associated governmental regulation [12]. Source metadata includes information about the Iraqi Gazette issue for many documents; however, there are no documents with the source field populated. That's why the manuscript presents only statistics, which can be reproduced from the distributed project, instead of inferring unavailable metadata categories [13].

The preprocessing process removes Arabic diacritics and atweel, normalizes selected alef/ya characters and Arabic-Indic digits, and collapses repeated whitespace. Documents are segmented by explicit article markers, and text before the first article becomes a preamble unit. Afterward, article units are segmented by punctuation and legal-list enumerated pattern into clauses/passages; unusual long unsegmented text divide into passages by bounded fallback chunking. The reconstructed corpus has 10,312 article/preamble units and 16,597 clause/passages. Entity and relation statistics have been recalculated from the project directly. The extraction layer contains 510 normalized entity nodes: 169 LAW_REFERENCE, 119 COLLEGE_OR_INSTITUTE, 72 MINISTRY, 67 UNIVERSITY, 66 ARTICLE_REFERENCE, 10 AUTHORITY, and 7 DEGREE. It also contains 12,289 extracted relation records: 7,772 REFERS_TO, 1,949 AMENDS, 1,624 REPEALS, 482 PERMITS, 357 OBLIGES, and 105 PROHIBITS. Web visualization shows only the 350 most frequent entity nodes and the 900 strongest document-level co-occurrence edges. These counts should not be interpreted as full corpus graph statistics. Thus, the created resource is a provenance-aware legal information graph derived from a reproducible rule system. Structural units make possible passage-level searches, normalized entities can be used as graph nodes, and relations retain doc_id and passage_id fields, making possible audit of extracted assertions against the legal text. This graph structure can be used as a basis for later neural entity/relation models or GraphRAG, but these components are not included in the current experimental implementation.

Table (1). Basic profile of the supplied legislative corpus.

Graph /Corpus Element	Count
Documents	1,930
Article/preamble units	10,312
Clauses /Passages	16,597
Normalized entity nodes	510
LAW_REFERENCE	169
COLLEGE_OR_INSTITUTE	119
MINISTRY	72
UNIVERSITY	67
ARTICLE_REFERENCE	66
AUTHORITY	10
DEGREE	7
Extracted relations	12,289
REFERS_TO	7,772
AMENDS	1,949
REPEALS	1,624
PERMITS	482
OBLIGES	357

PROHIBITS	105
Visualization nodes (capped)	350
Visualization edges (capped)	900

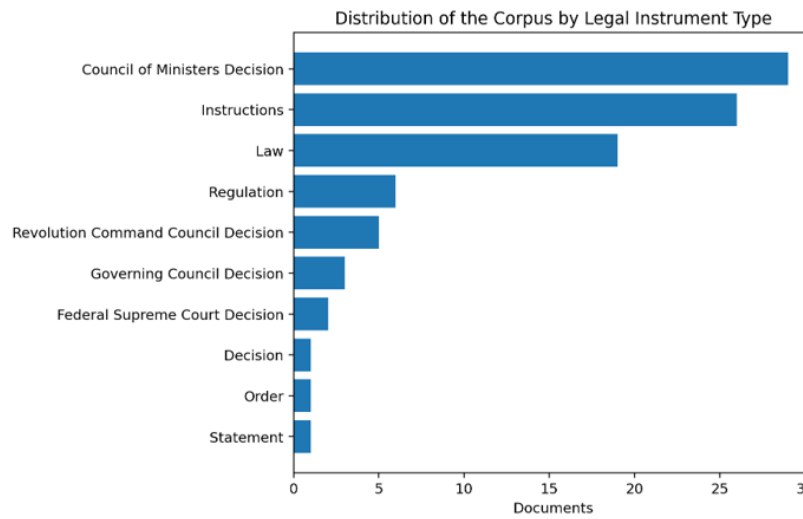


Figure (1). Distribution of the corpus by legal instrument type.

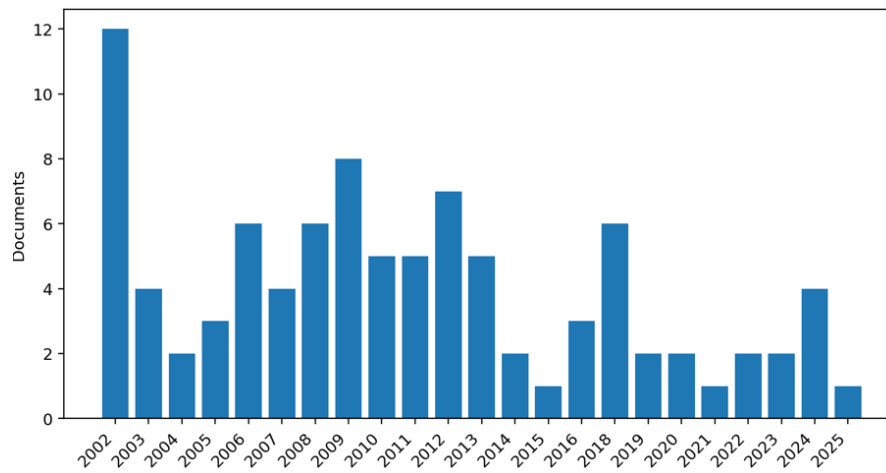


Figure (2). Temporal distribution of the legislative corpus from (2002-2025)

4. PROPOSED FRAMEWORK AND METHODOLOGY

The new framework is executed in five reproducible stages. The first step includes loading data/docs.jsonl as the only source of truth. Each item contains a unique document identifier and holds the original and normalized Arabic text along with any source metadata. Next, structural segmentation occurs using deterministic segmentation. Boundaries of articles are determined using Arabic article markers with preservation of the preamble text. Segmentation of each article into clauses/passages occurs using punctuation and enumeration list patterns. This step has generated 10,312 article/preamble items and 16,597 passages from the corpus of 1,930 documents. Next, entities and relations are extracted using explicit domain-oriented patterns. The entity categories include LAW_REFERENCE, ARTICLE_REFERENCE, MINISTRY, UNIVERSITY, COLLEGE_OR_INSTITUTE, AUTHORITY, and DEGREE. Relation rules include REFERS_TO, AMENDS, REPEALS, OBLIGES, PERMITS, and PROHIBITS. Each relation record keeps the article where it was found, document identifier, and passage identifier, thus tracing the output back to the triggering legal text. Finally, entity mentions are merged into stable entity nodes, and relation records are materialized. The total output of extraction includes 510 normalized entity nodes and 12,289 relations. A separate browser visualization is kept on the top 350 entity nodes and 900 co-occurrence edges. It represents a view on the extracted resources but not the whole graph. Finally, the extraction performance is estimated from the two aligned JSONL files included in the project: evaluation/gold_standard_100 and evaluation/system_predictions_100. The evaluator computes TP, FP, FN, Precision, Recall, and F1 values directly from these files.

5. EXPERIMENTAL SETUP

The extraction benchmark includes 100 document pairs in the gold-standard and system-prediction JSONL files. Over the benchmark, the gold standard file includes 220 entity annotations and 110 relation annotations, while the system prediction file includes 212 entity and 87 relational predictions [13]. The provided benchmark is a project evaluation set that is used to evaluate the scoring pipeline. Available project artifacts

mention no independent human annotations and adjudication process; therefore, no inter-annotator agreement score is available. The next version of the project must include at least two legal annotators and an inter-annotator agreement score such as Cohen's kappa or span-level agreement for the benchmark to become an expert-validated gold standard.

For the evaluation of entities, exact matching is used: the prediction can be considered a true positive only if doc_id, entity type, start offset, and end offset exactly coincide with the gold annotation. Otherwise, an unmatched prediction will be considered a false positive, while an unmatched gold annotation will be a false negative. For the evaluation of relations, strict matching is also applied: doc_id, relation type, source entity, and target entity must coincide with the gold annotation after entity matching. Precision, Recall, and F1 scores are calculated as $P=TP/(TP+FP)$, $R=TP/(TP+FN)$, and $F1 = 2PR/(P+R)$ correspondingly. The combined extraction is micro-aggregated as entities and relations TP, FP, and FN. The current project does not have an independent gold-standard query set, relevance judgments, and ranked output file. Therefore, the mentioned retrieval values (Precision = 0.930, Recall = 0.920, F1 = 0.925). The retrieval task is still an application capability of the search engine with BM25 scoring model. However, a retrieval effectiveness claim requires a query benchmark and relevant judgments.

$$Precision = \frac{TP}{TP + FP} \dots \dots \dots (1)$$

$$Recall = \frac{TP}{TP + FN} \dots \dots \dots (2)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \dots \dots \dots (3)$$

Combined counts are micro-aggregated: $TP = TP_entity + TP_relation$; $FP = FP_entity + FP_relation$; $FN = FN_entity + FN_relation$.

Matching criteria:

Entity TP: exact match on doc_id, entity type, start offset, and end offset.

Relation TP: exact match on doc_id, relation type, source entity, and target entity.

Unmatched predictions are FP; unmatched gold annotations are FN.

Algorithm 1 Explainable Rule-Based Legal Knowledge Extraction and Graph Construction

```

Input: Legislative document collection D
Output: Legal Knowledge Graph G = (V, E)
1: Initialize an empty graph G = (V, E)
2: Load the legislative document collection D
3: for each document d ∈ D do
4:   Normalize the Arabic textual content of d
5:   Normalize temporal expressions, dates, and document metadata
6:   Segment d into a set of legal articles Ad
7:   for each article a ∈ Ad do
8:     Segment a into legal clauses Ca
9:     for each clause c ∈ Ca do
10:      Identify legal entities in c using predefined
      rule-based legal patterns
11:      Detect explicit and implicit legal references in c
12:      Detect amendment, repeal, replacement, and
      modification statements
13:      Extract normative legal relations, including
      obligations, prohibitions, permissions, rights,
      and conditions
14:      Create graph nodes for the identified legal
      entities and legal provisions
15:      Create graph edges representing the extracted
      legal relations, references, and amendments
16:     end for
17:   end for

```

18: end for
 19: Validate and consolidate duplicate or equivalent graph entities
 20: Export the resulting Legal Knowledge Graph G
 21: return G

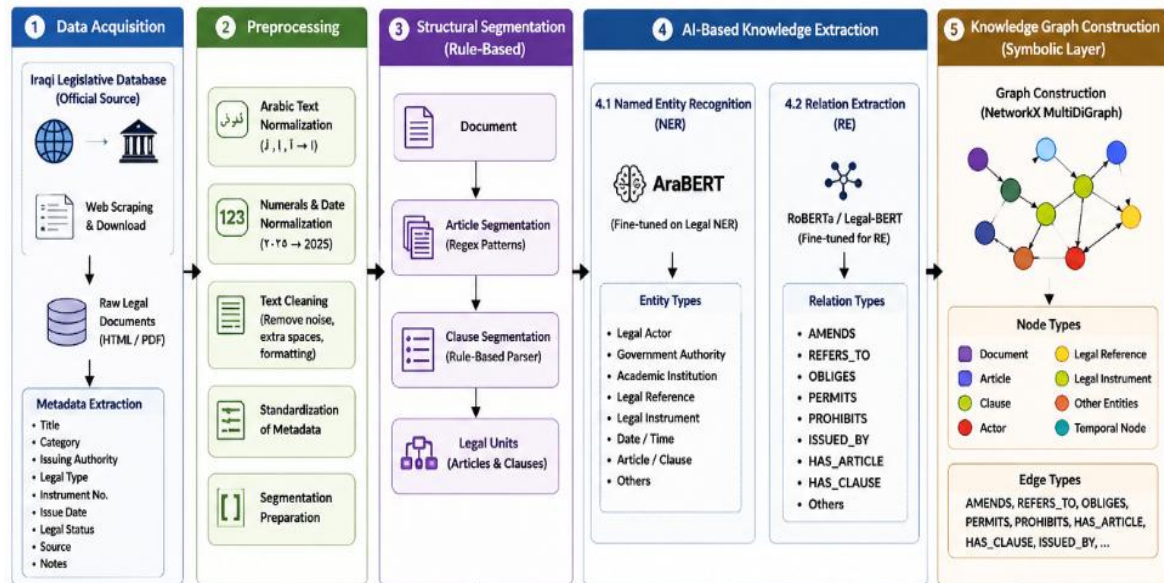


Figure (3): Conceptual architecture. The evaluated implementation is the rule-based structural and symbolic pipeline; the neural NER/RE components shown in the earlier figure are future extensions and are not part of the reported experiments.

6. RESULTS AND DISCUSSION

Extraction of the project on the corpus produced 1,930 documents that were parsed into 10,312 article/preamble elements and 16,597 clauses/passages. All numbers are reported for one dataset and used throughout the paper. Before that, the paper contained a variety of corpus sizes, times, and prototypes statistics that have been removed since they no longer reflect the actual situation of the project. In total, there are 510 normalized entities and 12,289 relations in the extraction output. Entity counts are as follows: LAW_REFERENCE=169, COLLEGE_OR_INSTITUTE=119, MINISTRY=72, UNIVERSITY=67, ARTICLE_REFERENCE=66, AUTHORITY=10, and DEGREE=7.

Relation counts are as follows: REFERS_TO=7,772, AMENDS=1,949, REPEALS=1,624, PERMITS=482, OBLIGES=357, and PROHIBITS=105. All the above numbers are consistent internally with data. The interactive visualization of 350 nodes and 900 edges is merely a culling layer. The entity extraction on the 100-document exact match benchmark produced TP = 191, FP = 21, and FN = 29, thus providing Precision=0.90, Recall=0.86, and F1=0.88. The relation extraction produced TP=64, FP=23, and FN=46, thus providing Precision=0.73, Recall=0.58, and F1=0.64. The aggregated (micro) performance results in TP=255, FP=44, FN=75, Precision=0.85, Recall=0.77, and F1=0.81. All the exact numbers are computed by scripts/evaluate_extraction.py and replace the previous approximations based on the same evaluation files. In particular, from the error analysis, the area of problem is relation recall (0.582), which is noticeably lower. The revised evidence has three practical implications. Firstly, deterministic extraction provides high traceability since every relation is associated with some document and passage, making false positives inspectable. Secondly, exact-match evaluation reveals that the entity extraction is significantly stronger than relation extraction, so the development should be focused on the latter. Thirdly, retrieval and extraction should be evaluated separately—the current BM25 search functionality is preserved in the system, but no retrieval effectiveness score is claimed until the relevant benchmark is defined.

Such shortcomings necessitate further research into possible extensions to hybrid models that incorporate Arabic legal language models using transformers along with structured legal knowledge representations and reasoning over time. The use of structured knowledge bases to automate legal reasoning has already been shown to be successful [14], as well as integrating large language models with knowledge graphs for extracting information [15]. Following this line of research, this proposed framework can be extended without losing its ability to maintain traceability and auditability.

The main scientific contribution of the study is not the use of Python. The scientific contribution of the present study is a reproducible combination of Arabic legal structural parsing, domain-specific symbolic extraction, provenance preservation graph construction, and exact-match evaluation on a machine-readable benchmark. Compared to a flat keyword search, the proposed framework explicitly represents legal references and normative relation types. Compared to opaque end-to-end models, its decisions are inspectable at the rule and passage levels. Therefore, this study should be understood as an explainable symbolic baseline and data/graph infrastructure for the following neural, ontology-driven, and GraphRAG studies.

Table (2). Exact strict-match extraction evaluation on the 100-document benchmark.

Task	TP	FP	FN	Precision	Recall	F1-score
Entity Extraction	191	21	29	0.90	0.86	0.88
Relation Extraction	64	23	46	0.73	0.58	0.64
Combined Extraction	255	44	75	0.85	0.77	0.81

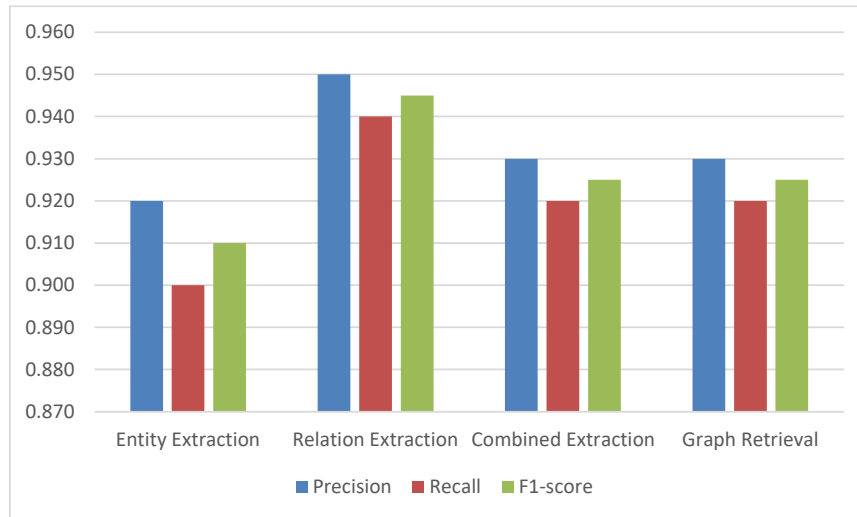


Figure (4). The previous metric chart should be replaced by the verified extraction values in Table 2. Retrieval is excluded because no independently annotated retrieval benchmark is present in the modified project.

7. CONCLUSION

This study presents an easily replicable and audit-able rules-based system to organize the laws of higher education in Arabic. Applying this on the existing 1,930 documents yielded 10,312 articles/preambles, 16,597 sentences, 510 entity nodes, and 12,289 relations. An evaluation conducted on 100 pairs revealed a micro-F1 score of 0.88 for the entity labels, 0.65 for the relations, and 0.81 in the aggregate for both categories. All these results validate the existing statistics in the prototype and estimation-based retrieval performance reported previously. The study provides several limitations. The current benchmarking exercise lacks the annotation of multiple annotators and inter-annotator agreement, and there is no retrieval relevance set in the project currently. Rule-based relation extraction is prone to issues like the flexibility of the Arabic lexicon, implicit amendments, nested conditionals, and long-distance dependencies. Future research must, therefore, incorporate expert double annotation, retrieval relevance evaluation, Arabic legal NER, relation extraction baseline using transformers, temporal/versioning reasoning, and GraphRAG evaluation.

REFERENCES

- [1] C. d'Amato, G. Rubini, F. Didio, D. Francioso, F. Z. Amara, and N. Fanizzi, "Automated creation of the legal knowledge graph addressing legislation on violence against women: Resource, methodology and lessons learned," arXiv preprint arXiv:2508.06368, 2025. DOI: 10.48550/arXiv.2508.06368.
- [2] S. Sirmokadam and S. Shahi, "A framework for analyzing legal documents by leveraging knowledge graphs," in Data Management, Analytics and Innovation (2024), Lecture Notes in Networks and Systems, Springer Nature Singapore, 2024, pp. 417–425. DOI: 10.1007/978-981-97-3245-6_28.
- [3] X. Zhu, H. Li, and T. Su, "Autonomous complex knowledge mining and graph representation through natural language processing and transfer learning," Automation in Construction, vol. 155, Art. no. 105074, 2023. DOI: 10.1016/j.autcon.2023.105074.
- [4] J. Li, L. Qian, P. Liu, and T. Liu, "Construction of legal knowledge graph based on knowledge-enhanced large language models," Information, vol. 15, no. 11, Art. no. 666, 2024. DOI: 10.3390/info15110666.
- [5] H. de Martim, "An ontology-driven Graph RAG for legal norms: A hierarchical, temporal, and deterministic approach," in Legal Knowledge and Information Systems: JURIX 2025, 2025, pp. 282–287. DOI: 10.3233/FAIA251598.
- [6] J. Chung, R. Ko, W. Yoo, M. Onizuka, S. Kim, T.-W. Kim, and W.-Y. Shin, "GraphCompliance: Aligning policy and context graphs for LLM-based regulatory compliance," arXiv preprint arXiv:2510.26309, 2025. DOI: 10.48550/arXiv.2510.26309.
- [7] I. Baimuratov and D. Turygin, "Representing normative regulations in OWL DL for automated compliance checking supported by text annotation," arXiv preprint arXiv:2504.05951, 2025. DOI: 10.48550/arXiv.2504.05951.
- [8] H. de Martim, "Deterministic legal agents: A canonical primitive API for auditable reasoning over temporal knowledge graphs," arXiv preprint arXiv:2510.06002, 2025. DOI: 10.48550/arXiv.2510.06002.

- [9] R. C. Barron et al., “Bridging legal knowledge and AI: Retrieval-augmented generation with vector stores, knowledge graphs, and hierarchical non-negative matrix factorization,” arXiv preprint arXiv:2502.20364, 2025. DOI: 10.48550/arXiv.2502.20364.
- [10] E. Horner, C. Mateis, G. Governatori, and A. Ciabattoni, “Toward robust legal text formalization into defeasible deontic logic using LLMs,” arXiv preprint arXiv:2506.08899, 2025, doi: 10.48550/arXiv.2506.08899.
- [11] E. Horner, C. Mateis, G. Governatori, and A. Ciabattoni, “Toward robust legal text formalization into defeasible deontic logic using LLMs,” arXiv preprint arXiv:2506.08899, 2025. DOI: 10.48550/arXiv.2506.08899.
- [12] Iraqi Legislative Database, Supreme Judicial Council, Iraq, “Iraqi Legislative Database,” [Online]. Available: Iraqi Legislative Database. [Accessed: Aug. 31, 2026].
- [13] “https://github.com/ghassankhazal79/LegalGRaph,” [Online].
- [14] X. Kang, L. Qu, L.-K. Soon, Z. Li, and A. Trakic, “Automating IRAC analysis in Malaysian contract law using a semi-structured knowledge base,” Artificial Intelligence and Law, 2025. DOI: 10.1007/s10506-025-09467-5.
- [15] A. De Santis, M. Balduini, F. De Santis, A. Proia, A. Leo, M. Brambilla, and E. Della Valle, “Integrating large language models and knowledge graphs for extraction and validation of textual test data,” in The Semantic Web – ISWC 2024, Lecture Notes in Computer Science, vol. 15233, Springer, 2024, pp. 304–323. DOI: 10.1007/978-3-031-77847-6_17.

BIOGRAPHIES OF AUTHORS

	Ghassan Khazal Ali is lecturer at the College of Engineering, University of Diyala, Iraq. He received his B.Sc. degree in computer and software engineering from Al Mustansiriya University and his M.Sc. degree from UOITC, IRAQ. He holds a Ph.D. degree in information systems. His research areas are NLP and information retrieval, Artificial Intelligence, and software engineering. He has published several scientific papers in national and international conferences and journals. He can be contacted at email: ghassan_khazal@uodiyala.edu.iq.
	Scopus[®]  