# ViHERMES: A Graph-Grounded Multihop Question Answering Benchmark and System for Vietnamese Healthcare Regulations

Long S. T. Nguyen, Quan M. Bui, Tin T. Ngo, Quynh T. N. Vo, Dung N. H. Le, and Tho T. Quan(✉)

URA Research Group, Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), VNU-HCM, Ho Chi Minh City, Vietnam
qttho@hcmut.edu.vn

**Abstract.** Question Answering (QA) over regulatory documents is inherently challenging due to the need for multihop reasoning across legally interdependent texts, a requirement that is particularly pronounced in the healthcare domain where regulations are hierarchically structured and frequently revised through amendments and cross-references. Despite recent progress in retrieval-augmented and graph-based QA methods, systematic evaluation in this setting remains limited, especially for low-resource languages such as Vietnamese, due to the lack of benchmark datasets that explicitly support multihop reasoning over healthcare regulations. In this work, we introduce the Vietnamese Healthcare Regulations–Multihop Reasoning Dataset (ViHERMES), a benchmark designed for multihop QA over Vietnamese healthcare regulatory documents. ViHERMES consists of high-quality question–answer pairs that require reasoning across multiple regulations and capture diverse dependency patterns, including amendment tracing, cross-document comparison, and procedural synthesis. To construct the dataset, we propose a controlled multihop QA generation pipeline based on semantic clustering and graph-inspired data mining, followed by large language model–based generation with structured evidence and reasoning annotations. We further present a graph-aware retrieval framework that models formal legal relations at the level of legal units and supports principled context expansion for legally valid and coherent answers. Experimental results demonstrate that ViHERMES provides a challenging benchmark for evaluating multihop regulatory QA systems and that the proposed graph-aware approach consistently outperforms strong retrieval-based baselines. The ViHERMES dataset and system implementation are publicly available at https://github.com/ura-hcmut/ViHERMES.

**Keywords:** Multihop Question Answering · Healthcare Regulations · Graph-Aware Retrieval · Low-Resource Languages

L. S. T. Nguyen and Q. M. Bui—Equal contribution.


N. T. Nguyen et al. (Eds.): ACIIDS 2026, LNAI 16530, pp. 321–336, 2026.
https://doi.org/10.1007/978-981-92-0071-9_22

## 1 Introduction

Accessing and complying with regulatory documents is a critical requirement in modern societies, particularly in domains where legal correctness directly affects public safety and service quality. As governments increasingly publish regulations in digital form, there is a growing demand for intelligent *Question Answering* (QA) systems that can assist users in understanding and navigating complex legal and administrative texts [12].

The healthcare domain presents a particularly challenging setting for regulatory QA. Unlike medical articles or clinical guidelines, which are often descriptive and relatively self-contained, healthcare regulations are legal documents with strict administrative structures and dense inter-document dependencies. These documents are typically organized in a hierarchical manner, with clearly defined units such as articles and clauses, and are frequently interconnected through legal relations such as amendments, replacements, supplements, and cross-references. As a result, regulatory QA in healthcare is inherently a multihop problem rather than a single-document comprehension task. In practice, multihop reasoning becomes a fundamental requirement, as correct answers may depend on identifying the latest effective provision, tracing chains of amendments across multiple documents, and combining definitions or procedural rules referenced from separate regulations, rendering single-hop retrieval or isolated passage matching fundamentally insufficient in this setting [8,14].

Recent advances in Artificial Intelligence, particularly in Natural Language Processing, have significantly improved the capabilities of QA systems. *Retrieval-Augmented Generation* (RAG) frameworks have become a dominant paradigm by combining neural retrieval models with *Large Language Models* (LLMs) to generate answers grounded in retrieved contexts [9]. While naive RAG approaches enhance factual coverage, they typically treat documents as flat collections of passages and lack explicit awareness of legal structures and inter-document relations [16]. In regulatory domains, this limitation often leads to answers that are linguistically plausible but legally incomplete or outdated. To address these shortcomings, graph-based retrieval and data mining methods such as GraphRAG [3], HippoRAG [6], and LightRAG [5] have been proposed to incorporate structured knowledge into the retrieval process. However, when applied to regulatory documents, such automatically induced graphs often struggle to faithfully capture the formal and hierarchical nature of legal texts [11]. In particular, multihop regulatory questions that depend on amendment chains or legal validity require precise, rule-aware relationships that are difficult to recover from generic entity–relation extraction alone.

Vietnamese is a low-resource language, and datasets that jointly address the healthcare and legal or regulatory domains remain particularly scarce. Despite the practical importance of multihop regulatory QA, the lack of high-quality benchmark datasets has hindered systematic research in this area. Existing Vietnamese legal QA datasets do include regulatory texts; however, they primarily focus on general legal documents or regulations in other domains such as education, and are typically designed for single-document comprehension or shallow

QA rather than multihop reasoning across documents [1,2,7,13,17]. Similarly, existing Vietnamese healthcare QA datasets largely concentrate on clinical narratives, medical records, or healthcare-related news articles, rather than healthcare regulatory documents [15,19,21]. Crucially, none of these datasets explicitly support multihop reasoning over legally interdependent healthcare regulations. As a result, the performance and analytical understanding of QA systems in this domain remain difficult to evaluate in a controlled and meaningful manner.

In this work, we introduce ***Vi****etnamese* ***HE****althcare* ***R****egulations–****M****ultihop* *R***E***asoning Data***S***et* (ViHERMES), the first benchmark dataset specifically designed for multihop QA over Vietnamese healthcare regulations. ViHERMES focuses on questions that inherently require reasoning across multiple regulatory documents, capturing diverse dependency patterns such as amendment tracing, cross-document comparison, and procedural synthesis. To construct the dataset, we propose a controlled multihop QA generation pipeline that leverages semantic clustering as a form of graph-inspired data mining to sample coherent sets of regulatory contexts, followed by LLM-based QA generation with structured evidence and reasoning annotations. Beyond the dataset, we also propose a graph-aware retrieval framework tailored to the hierarchical and legally constrained nature of regulatory documents. Our approach represents regulations at the level of legal units and explicitly models formal legal relations, enabling principled context expansion strategies that respect legal validity and avoid context drift. This retrieval mechanism is integrated into a multi-agent QA system that separates intent understanding, graph-based retrieval, and answer verification, thereby supporting robust and intelligent healthcare regulatory QA. Our main contributions are summarized as follows.

- We introduce ViHERMES, the first benchmark dataset for Vietnamese healthcare regulatory QA that explicitly targets multihop reasoning across legally interdependent documents.
- We propose a graph-inspired data mining pipeline for controlled multihop QA dataset construction with high-quality evidence-grounded annotations.
- We present a graph-aware, multi-agent QA framework that effectively leverages legal structure and demonstrates consistent improvements over strong retrieval-based baselines on ViHERMES.

## 2 Related Works

### 2.1 Multihop QA and Graph-Based Retrieval

Multihop QA has been extensively studied to address queries that require reasoning over multiple pieces of evidence rather than a single passage. Early approaches such as IRCoT [20] interleave retrieval with chain-of-thought reasoning to iteratively guide evidence selection, while RAPTOR [18] organizes textual units into hierarchical summaries that support retrieval across different levels of abstraction. Building upon these ideas, recent graph-based retrieval frameworks, including GraphRAG [3], HippoRAG [6], LightRAG [5], and MiniRAG

[4], incorporate graph structures into retrieval-augmented generation by modeling entities, text chunks, or their relations as nodes and enabling neighborhood expansion or graph-based ranking. These methods have demonstrated strong performance gains on general multihop QA benchmarks and improved efficiency, particularly in resource-constrained settings. However, most existing approaches are primarily designed for unstructured or loosely structured text collections and rely on automatically induced, largely entity-centric graphs. Consequently, they struggle to faithfully capture the formal hierarchy, rule-based dependencies, and legal validity constraints inherent in regulatory documents. This limitation is especially pronounced in the healthcare domain, where answering a question often requires tracing amendment chains, identifying currently effective provisions, and combining legally interdependent clauses across multiple documents. In contrast, our work explicitly aligns both dataset construction and retrieval with the intrinsic legal structure of healthcare regulations, enabling principled multihop reasoning over legally interdependent documents.

### 2.2 Vietnamese Regulation-Related QA Datasets

Benchmark datasets for Vietnamese QA remain relatively scarce, particularly for domains involving structured regulatory documents. Existing datasets that cover regulations are primarily situated in the legal or educational domains. ALQAC [1] introduces a manually annotated legal QA dataset based on Vietnamese statute laws, while several other works focus on regulations in specific domains, including educational management [13], bidding law (ViBidLQA) [7], university training regulations (ViRHE4QA) [2], and large-scale labor law retrieval with associated QA data [17]. Although these datasets provide valuable resources for Vietnamese legal QA, they are typically designed for single-document comprehension or shallow question answering and do not explicitly support multihop reasoning across legally interdependent regulations. In contrast, existing Vietnamese healthcare QA datasets largely focus on clinical or informational content rather than regulatory texts. ViMedAQA [19] targets abstractive medical QA over clinical topics such as diseases and drugs, UIT-ViNewsQA [21] is constructed from healthcare news articles for machine reading comprehension, and ViHealthQA [15] collects expert-answered questions from health websites. While these datasets are valuable for healthcare-related QA, they do not reflect the hierarchical structure, amendment relationships, or cross-document dependencies that characterize healthcare regulations. Consequently, none of the existing Vietnamese datasets jointly address healthcare regulations and multihop reasoning over legally interdependent documents, leaving a critical gap that our ViHERMES dataset is specifically designed to fill.

## 3 ViHERMES Dataset

### 3.1 Dataset Creation

We construct ViHERMES through a controlled dataset creation pipeline that integrates human supervision, semantic clustering, and LLM-based generation to ensure genuine multihop reasoning and evidence-grounded answers over health-care regulatory documents. An overview of the pipeline is shown in Fig. 1.

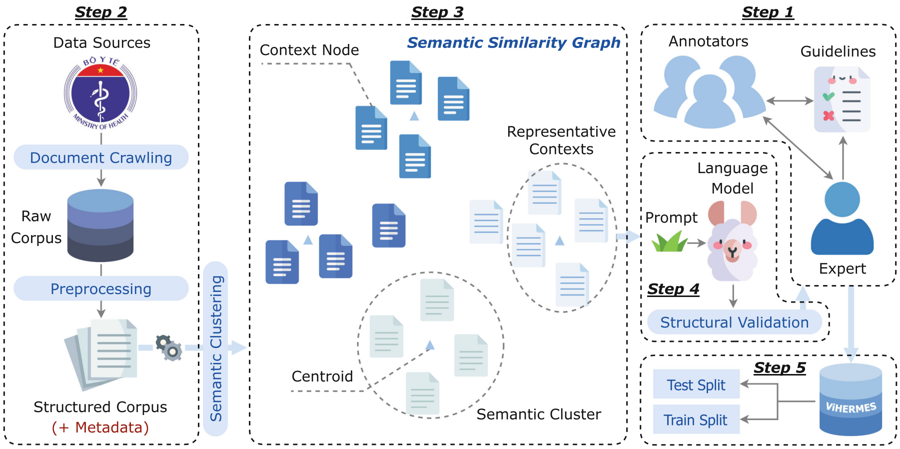


**Fig. 1.** ViHERMES dataset construction pipeline, from corpus collection and semantic clustering to LLM-based generation and splitting.

**Step 1: Annotator Recruitment and Guideline Design.** We recruit a small team of *annotators* together with a domain *expert*. The expert provides high-level guidance to ensure multihop validity and legal consistency. The annotation process follows shared *guidelines* defining multihop criteria, evidence usage, and legal validity constraints.

**Step 2: Corpus Collection and Preprocessing.** Regulatory documents are collected from official *data sources*, such as public portals of the Vietnamese Ministry of Health[1], via automated *document crawling*. The collected texts form a *raw corpus*, which is subsequently processed by a *preprocessing* pipeline that normalizes content, removes noise, and segments documents according to their legal hierarchy. The resulting *structured corpus*, together with associated *metadata*, consists of context units corresponding to legal units (e.g., articles or clauses), each represented as a context triple $\langle \texttt{id}, \texttt{title}, \texttt{text} \rangle$.

**Step 3: Semantic Clustering and Similarity Graph Induction.** Each context is treated as a *context node* and encoded into an embedding vector $\mathbf{e}_i \in \mathbb{R}^d$ using a neural embedding model. To emphasize semantic direction rather than vector magnitude, embeddings are $\ell_2$-normalized as $\tilde{\mathbf{e}}_i = \mathbf{e}_i / \|\mathbf{e}_i\|_2$. We then perform *semantic clustering* using $K$-Means on the normalized vectors, which corresponds to *SphericalK-Means* in practice. Each resulting *semantic cluster* $C_k$ is summarized by its *centroid* $\mathbf{c}_k = \frac{1}{|C_k|} \sum_{\tilde{\mathbf{e}}_i \in C_k} \tilde{\mathbf{e}}_i$. Contexts are ranked

[1] https://moh.gov.vn/web/ministry-of-health.

by their Euclidean distance to the centroid $d(\tilde{\mathbf{e}}_i, \mathbf{c}_k) = \|\tilde{\mathbf{e}}_i - \mathbf{c}_k\|_2$, and the nearest ones are selected as *representative contexts*. This cluster-and-centroid organization induces an implicit *semantic similarity graph* that preserves topical coherence among regulatory texts. The graph is not an explicit legal graph; rather, it provides a latent similarity structure without relying on potentially noisy automatic extraction of legal relations.

**Step 4: Multihop QA Generation and Structural Validation.** For a given hop level $h$, we sample tuples of contexts $\mathcal{T} = \{c_1, \ldots, c_h\}$ from the pool of *representative contexts* under two constraints. First, contexts within a tuple must originate from different document `titles`, i.e., $\forall i \neq j$, $\texttt{title}(c_i) \neq \texttt{title}(c_j)$, to ensure cross-document reasoning. Second, to maintain diversity, the overlap between any two tuples $\mathcal{T}_a$ and $\mathcal{T}_b$ is bounded by $|\mathcal{T}_a \cap \mathcal{T}_b|/|\mathcal{T}_a| \leq \tau$. Each tuple is then provided to a *language model* through a structured *prompt* to generate multihop question–answer pairs, where each answer must rely on all $h$ contexts. In practice, each hop corresponds to one legal dependency step, such as amendment tracing, definition lookup, or cross-document procedural composition. The generated outputs undergo *structural validation*, which verifies format correctness, consistency between hop level and evidence usage, and explicit reasoning. Ambiguous or borderline cases are further reviewed by annotators under expert supervision to ensure legal coherence and correctness.

**Step 5: Dataset Aggregation and Splitting.** After validation, the accepted question–answer instances are aggregated and exported in a structured format. The dataset is then divided into a *train split* and a *test split*, with controlled distributions over hop levels, forming the final *ViHERMES* benchmark.

### 3.2 Dataset Statistics

Table 1 presents a representative instance from the ViHERMES dataset, including the query, answer, number of reasoning hops, explicit reasoning, supporting evidence, and corresponding context identifiers. Each identifier maps to a regulatory context stored as a triple $\langle \texttt{id}, \texttt{title}, \texttt{text} \rangle$. The `reasoning` and `evidence` fields provide explicit explanations and extractive textual spans grounded in the referenced contexts, respectively. Both fields are generated during dataset construction (Step 4) and subsequently verified by annotators, improving transparency and answer reliability. Table 2 summarizes key statistics of the ViHERMES test set across different *Numbers of Hops* (NoH). For each hop level, we report the number of samples, vocabulary size, and length statistics of questions, answers, and supporting contexts. As NoH increases, question and answer lengths grow substantially, reflecting higher reasoning complexity, while context lengths expand rapidly, indicating the need to integrate information from long and diverse regulatory texts.

## 4 Graph-Aware System for Regulatory QA

Figure 2 illustrates the proposed *graph-aware QA system* for Vietnamese healthcare regulations. Unlike standard RAG pipelines that treat the corpus as a flat

Table 1: A representative multihop QA instance from the ViHERMES dataset.

| **Field** | **Content** |
|---|---|
| Query | **English:** To ensure comprehensive legal compliance, what main obligations must a healthcare facility fulfill regarding service pricing during the transition period, maintenance of baseline quality, and the legal conditions required for lawful operation? |
| Answer | **English:** To fully comply with legal requirements, a healthcare facility must fulfill the following obligations: (1) Operational conditions: the facility must obtain an operating license, and all practicing physicians must hold valid professional certificates. (2) Baseline quality: the facility must implement quality standards, conduct reviews and corrective actions, and perform self-assessments with mandatory reporting as required. (3) Service pricing during the transition period: when a new pricing decision is issued, patients who began treatment earlier must continue to be charged the previous price until discharge or completion of outpatient treatment. |
| Number of hops | 3 |
| Context IDs | [4814–7506-4302, 7115-6545-3240, 6442-4111-1890] |
| Evidence | **English:** The evidence consists of three regulatory provisions covering service pricing during transitional periods, baseline quality responsibilities, and legal requirements for operation, respectively. |
| Reasoning | **English:English:** The answer integrates three regulatory sources covering operational legality, baseline quality, and transitional service pricing, forming a legally consistent response. |

Table 2: Statistics of the ViHERMES test split across different multihop settings. Lengths are reported as *min* / *mean* / *max*.

| NoH | NoS | Vocab | Question length | Answer length | Context length |
|---|---|---|---|---|---|
| 1 | 312 | 2232 | 48 / 113.2 / 215 | 58 / 181.8 / 504 | 195 / 1185.1 / 4257 |
| 2 | 312 | 2949 | 120 / 249.4 / 446 | 89 / 399.1 / 810 | 734 / 2729.4 / 10075 |
| 3 | 312 | 3178 | 168 / 325.7 / 688 | 135 / 524.0 / 1100 | 743 / 3739.8 / 14767 |
| 4 | 312 | 3545 | 179 / 386.4 / 813 | 175 / 748.0 / 1696 | 1461 / 5677.5 / 16556 |
| 5 | 312 | 3707 | 194 / 445.9 / 977 | 101 / 889.9 / 1938 | 1115 / 6451.6 / 18878 |
| All | 1560 | 5703 | 48 / 304.2 / 977 | 58 / 548.8 / 1938 | 195 / 3953.9 / 18878 |

collection of passages, our system operates over a *structure-driven regulatory knowledge graph*, where legal texts are modeled as interconnected *regulatory nodes* linked by defined relations. The system follows a *seeded retrieval and propagation* paradigm, retrieving legally valid and contextually complete evidence.

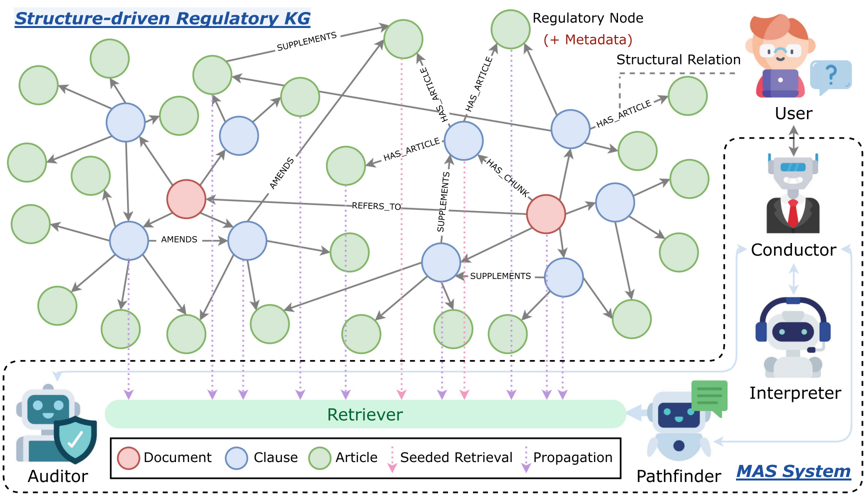


**Fig. 2.** Overview of the structure-driven regulatory QA system.

### 4.1 Structure-Driven Regulatory Knowledge Graph

The system is built upon a *Structure-driven Regulatory Knowledge Graph* (SRKG), which encodes regulatory documents according to formal legal drafting conventions. Unlike entity-centric or LLM-induced graphs, SRKG is *structure-driven*, with nodes and edges corresponding to legally grounded units and relations. The graph is organized as a *flat regulatory base layer*, where legal-unit nodes coexist on a unified plane and are retrieved directly, rather than through a top-down pipeline. Formally, SRKG is defined as a directed labeled graph $G = (V, E)$, where $V$ denotes regulatory nodes and $E \subseteq V \times \mathcal{R} \times V$ represents typed edges with labels from $\mathcal{R} = \mathcal{R}_{\text{struct}} \cup \mathcal{R}_{\text{legal}}$.

**Regulatory Nodes.** SRKG represents regulatory content using *regulatory nodes* augmented with provenance metadata. As illustrated in Fig. 2, we distinguish three conceptual levels: *Document nodes* representing legal documents, *Clause nodes* corresponding to legal units with standalone meaning, and *Article nodes* capturing finer-grained units where applicable. Each node $v \in V$ is represented as $v = \langle \texttt{id}(v), \texttt{text}(v), \texttt{meta}(v) \rangle$, where $\texttt{meta}(v)$ includes document identifier, promulgation date, issuing authority, and legal status. Each node is assigned a deterministic identifier following a strict scheme (e.g., `DocID::UnitID`), enabling stable referencing, amendment-chain tracing, and incremental ingestion.

**Relations.** SRKG encodes two complementary categories of relations, namely *structural relations* and *legal semantic relations*. Structural relations $\mathcal{R}_{\text{struct}} = \{\textsc{Has\_Article}, \textsc{Has\_Clause}\}$ preserve legal hierarchy and provenance. Legal semantic relations $\mathcal{R}_{\text{legal}} = \{\textsc{Amends}, \textsc{Replaces}, \textsc{Supplements}, \textsc{Refers\_To}\}$ capture inter-document dependencies affecting legal interpretation and validity. All relations are induced using rule-based pattern matching over standardized legal expressions (e.g.,

[refers to Article Z]), yielding high precision and avoiding spurious links. When a regulatory node references a legal unit not yet present in the corpus, the system creates a lightweight *placeholder node* to preserve the structural dependency. Such placeholders are automatically resolved when the corresponding documents are later ingested, ensuring graph completeness without disrupting downstream retrieval.

### 4.2 Seeded Retrieval and Propagation

Given a user query $q$, the system performs retrieval through a two-stage *seeded retrieval and propagation* strategy.

**Seeded retrieval.** Retrieval is first conducted over the legal-unit base layer. The *Pathfinder* identifies a small set of highly relevant *seed regulatory nodes* $S_K(q) = \{v_1, \ldots, v_K\} \subseteq V$ by ranking nodes using a hybrid relevance score $s(v, q) = \lambda\, s_{\text{dense}}(v, q) + (1 - \lambda)\, s_{\text{sparse}}(v, q)$, where $s_{\text{dense}}$ measures semantic similarity in embedding space and $s_{\text{sparse}}$ captures lexical evidence (e.g., BM25). The resulting $S_K(q)$ serves as anchor points for downstream expansion.

**Propagation.** Starting from $S_K(q)$, the system expands context in a controlled manner along relation-specific edges. For a node $v$ and relation $r$, outgoing neighbors are denoted as $\mathcal{N}_r(v) = \{u \mid (v, r, u) \in E\}$. This process activates three principled flows:

- **Validity tracing.** For $r \in \{\textsc{Amends}, \textsc{Replaces}, \textsc{Supplements}\}$, amendment chains are recursively traversed until a terminal node is reached. A node $u$ is terminal if $\forall r' \in \{\textsc{Amends}, \textsc{Replaces}, \textsc{Supplements}\}$, $\mathcal{N}_{r'}(u) = \emptyset$, and such terminal nodes are preferred as legally applicable evidence.
- **Contextual supplementation.** For reference relations ($r = \textsc{Refers\_To}$), expansion is restricted to direct neighbors only, providing essential legal context while preventing uncontrolled drift.
- **Provenance retrieval.** Structural relations are used to retrieve upstream document metadata, ensuring provenance-aware evidence usage.

**Context Assembly.** The final retrieval output is assembled as a bounded context set $C(q) = S_K(q) \cup P(q)$, where $P(q)$ denotes nodes obtained via relation-aware expansion. Propagation depth and relation types are explicitly constrained before $C(q)$ is passed to downstream agents.

### 4.3 Multi-agent System

The QA pipeline is implemented as a lightweight *Multi-Agent System* (MAS) composed of functionally specialized agents, as illustrated in Fig. 2. Each agent fulfills a distinct operational role within a unified regulatory QA workflow.

**Interpreter.** The *Interpreter* performs intent analysis and routing. It determines whether a query is regulatory, extracts key signals such as document identifiers, and decides whether graph-based retrieval is required. This step avoids unnecessary retrieval, improving overall execution efficiency.

**Pathfinder.** The *Pathfinder* implements the core graph-aware retrieval logic. It performs seeded retrieval over regulatory nodes followed by relation-aware propagation on the SRKG, including validity tracing along amendment relations and bounded expansion over reference relations. The agent returns a legally coherent and up-to-date context set $C(q)$.

**Auditor.** The *Auditor* verifies the retrieved evidence and intermediate outputs by checking consistency between regulatory nodes, relations, and generated claims. It performs grounding validation and safety checks to detect unsupported or potentially hallucinated content.

**Conductor.** The *Conductor* orchestrates the overall QA process. Based on the verified context $C(q)$, it invokes an LLM to generate the final natural-language answer and coordinates interactions among agents. When insufficient grounding is detected by the *Auditor*, the *Conductor* enforces conservative behavior such as abstention or clarification requests.

Overall, the proposed system integrates *structure-driven regulatory representation*, *seeded retrieval*, and *relation-aware propagation* within a unified multi-agent framework. By operationalizing legal structure directly during retrieval, the system enables legally valid multihop reasoning and robust evidence grounding, as demonstrated in our experiments.

## 5 Experimentations

We assess system performance along multiple complementary dimensions, including answer correctness, multihop reasoning quality, inference latency, graph construction overhead, and the effectiveness of agent coordination, and compare against representative baseline approaches.

### 5.1 Dataset

All experiments are conducted on the test split of ViHERMES, which is specifically designed to assess multihop reasoning over Vietnamese healthcare regulatory documents. The test set spans diverse hop levels and legal reasoning patterns, including amendment tracing, cross-document dependency resolution, and validity-aware clause composition.

### 5.2 Baselines

We compare the proposed system against representative baselines drawn from three major families of retrieval-augmented and multihop QA approaches.

**Naive RAG.** We consider standard RAG pipelines based on flat passage retrieval, including (1) lexical retrieval using BM25, (2) dense retrieval via embedding similarity, and (3) hybrid retrieval combining lexical and dense scores through weighted interpolation. These baselines represent common RAG settings without explicit multihop or structural awareness.

**Reasoning-guided Multihop QA.** IRCoT [20] is a multi-step QA approach that interleaves retrieval with *Chain-of-Thought* (CoT) reasoning to iteratively guide evidence selection using a strong LLM, without explicitly modeling document structure or legal validity constraints.

**Graph-based RAG Methods.** We include several representative graph-aware retrieval frameworks that incorporate graph structures into indexing and retrieval to support multihop reasoning:

- **RAPTOR** [18] constructs a hierarchical tree of textual units via recursive summarization, enabling retrieval across multiple levels of abstraction.
- **MiniRAG** [4] is a lightweight graph-based system that combines a small language model with a heterogeneous graph index to achieve efficient structured retrieval under limited computational budgets.
- **LightRAG** [5] emphasizes efficient graph construction and traversal strategies to balance retrieval quality and inference latency.
- **HippoRAG 2** [6] draws inspiration from hippocampal memory mechanisms, modeling long-term knowledge as a graph to improve recall and evidence integration across distant reasoning hops.

### 5.3 Evaluation Metrics

We evaluate system performance at both the answer and retrieval levels. For answer quality, we report token-level $F1$ to measure surface-level overlap between predicted and reference answers. Since $F1$ alone is insufficient to assess reasoning correctness and legal validity in multihop regulatory QA, we additionally adopt an *LLM-as-a-Judge* evaluation protocol [10], where a strong external LLM evaluates answers in terms of correctness, completeness, and consistency with supporting evidence. To assess retrieval quality independently of answer generation, we use Recall@$k = \frac{|E_{\text{gold}} \cap E^{(k)}_{\text{retrieved}}|}{|E_{\text{gold}}|}$, which measures the proportion of gold supporting contexts recovered within the top-$k$ retrieved results and reflects the effectiveness of evidence retrieval for multihop reasoning.

### 5.4 Experimental Setup

To ensure fair comparison and reproducibility, all systems share the same language model backbone and evaluation protocol. We adopt `GPT-4o-mini`[2] as the unified LLM backbone for answer generation, chosen for its balance between reasoning capability and computational efficiency. Text embeddings are generated using OpenAI `text-embedding-3-small`[3], a multilingual embedding model supporting both Vietnamese and English. All baselines are evaluated using default hyperparameters to reflect realistic out-of-the-box performance. LLM-as-a-Judge

Table 3: Main QA performance on the ViHERMES test set.

| Method | **F1** | LLM Judge | **Recall@5** |
|---|---|---|---|
| Naive RAG (BM25) | 0.3076 | 0.2027 | 0.2617 |
| Naive RAG (Dense) | 0.3289 | 0.2433 | 0.3241 |
| Naive RAG (Hybrid) | 0.4127 | 0.3324 | 0.3989 |
| IRCoT | 0.4835 | 0.3751 | 0.4254 |
| MiniRAG | 0.5429 | 0.4856 | 0.5083 |
| RAPTOR | 0.5941 | 0.5783 | 0.5563 |
| LightRAG | 0.7855 | 0.6756 | 0.7256 |
| HippoRAG 2 | 0.8023 | 0.7332 | 0.8032 |
| Ours | **0.8334** | **0.7554** | **0.8461** |
| w/o Auditor | 0.8150 | 0.6823 | 0.8267 |
| w/o Interpreter | 0.6540 | 0.5434 | 0.6134 |
| w/o Pathfinder[a] | 0.7734 | 0.6927 | 0.7955 |

[a] Instead of SRKG-based seeded retrieval and relation-aware propagation, *Pathfinder* is replaced with standard dense–sparse retrieval over flat text units.

[2] https://platform.openai.com/docs/models/gpt-4o-mini.
[3] https://platform.openai.com/docs/models/text-embedding-3-small.

evaluations are conducted using the `GPT-4o`[4] API under a standardized rubric to ensure consistent and comparable scoring across systems.

### 5.5 Results and Analysis

Table 3 reports the QA performance of all evaluated systems on the ViHERMES test set, covering answer quality ($F1$), reasoning reliability (LLM-as-a-Judge), and retrieval effectiveness (Recall@5). The results show a clear progression from flat, structure-agnostic RAG pipelines to multihop and graph-based approaches, indicating that Vietnamese healthcare regulatory QA requires structured multi-hop evidence integration. Among naive RAG baselines, Dense retrieval improves over BM25, and Hybrid retrieval further benefits from combining semantic and lexical signals; however, these approaches remain behind multihop methods due to the lack of explicit modeling of inter-document dependencies and legal validity constraints. IRCoT improves both $F1$ and LLM Judge scores through iterative reasoning, but remains limited by operating over unstructured text without explicit regulatory awareness. Graph-based methods achieve markedly stronger results: MiniRAG and RAPTOR benefit from graph structure, while LightRAG and HippoRAG 2 further improve answer quality and Recall@5, reflecting more effective multihop evidence integration. Overall, the proposed system achieves the best performance across all metrics, improving $F1$ by 3.1 points over HippoRAG 2 and confirming the effectiveness of structure-driven legal modeling with relation-aware propagation.

Ablation results further clarify component contributions. Removing the *Interpreter* causes the largest degradation ($F1$: 0.8334 → 0.6540), highlighting the importance of intent analysis and query routing. Excluding the *Auditor* leads to reduced reasoning reliability (LLM Judge: 0.7554 → 0.6823), underscoring

Table 4: Inference efficiency and graph construction statistics.

| Method | Inference efficiency | | Graph construction | | |
|---|---|---|---|---|---|
| | Avg. latency (s) | Avg. tokens | Nodes | Edges | Graph tokens |
| Naive RAG (BM25) | 4.1139 | 3009.8473 | – | – | – |
| Naive RAG (Dense) | 6.1139 | 3043.3433 | – | – | – |
| Naive RAG (Hybrid) | 9.2348 | 3289.3246 | – | – | – |
| IRCoT | 11.8312 | 4457.4561 | – | – | – |
| MiniRAG | 13.1923 | 4224.6595 | 2988 | 4323 | 3,270,523 |
| RAPTOR | 14.2325 | 4320.5432 | 3352 | 7234 | 3,436,236 |
| LightRAG | 17.3236 | 4558.3425 | 3569 | 9332 | 3,684,343 |
| HippoRAG 2 | 22.0278 | 4859.3696 | 2988 | 11183 | 3,966,963 |
| Ours | 14.7415 | 4236.4620 | 3727 | 12540 | 3,397,126 |

[4] https://platform.openai.com/docs/models/gpt-4o.

the role of guardrail verification. Replacing the *Pathfinder* with flat hybrid retrieval also degrades performance, confirming the benefit of seeded retrieval with relation-aware propagation over the SRKG.

Table 4 summarizes inference efficiency and graph construction. Compared with Naive RAG and IRCoT, graph-based approaches incur higher overhead from indexing and traversal; however, explicit legal modeling and constrained propagation achieve a favorable effectiveness–efficiency trade-off. Despite constructing a larger SRKG, the proposed system maintains competitive inference latency (14.74 s on average), comparable to RAPTOR (14.23 s) and faster than HippoRAG 2. Moreover, it outperforms LightRAG and HippoRAG 2 while using fewer graph tokens, indicating efficient context construction.

## 6 Conclusion

This paper introduces ViHERMES, a benchmark for multihop QA over Vietnamese healthcare regulations that captures the legally interdependent and frequently amended nature of regulatory texts. ViHERMES provides evidence-grounded question–answer pairs with explicit reasoning annotations across diverse dependency patterns, enabling systematic evaluation of multihop regulatory QA in a low-resource language setting. To support this benchmark, we propose a SRKG and a graph-aware, multi-agent QA system that combines seeded retrieval, relation-aware propagation, and verification, achieving consistent gains over strong RAG and graph-based baselines while maintaining competitive inference latency and efficient context construction. Beyond healthcare, the proposed structure-driven formulation is applicable to other regulatory domains with hierarchical organization and formal cross-document dependencies, and future work will focus on constraining propagation volume and enriching temporal validity for improved robustness in practical deployment.

## Bibliography


1. Do, D.T., et al.: A summary of the ALQAC 2024 competition. In: 2024 16th International Conference on Knowledge and System Engineering (KSE), pp. 422–427 (2024). https://doi.org/10.1109/KSE63888.2024.11063484
2. Do, T.P.P., Cao, N.D.D., Tran, K.Q., Van Nguyen, K.: R2GQA: retriever-reader-generator question answering system to support students understanding legal regulations in higher education. Artif. Intell. Law (2025). https://doi.org/10.1007/s10506-025-09457-7
3. Edge, D., et al.: From local to global: a graph rag approach to query-focused summarization (2024)
4. Fan, T., Wang, J., Ren, X., Huang, C.: MiniRAG: towards extremely simple retrieval-augmented generation. arXiv preprint arXiv:2501.06713 (2025)
5. Guo, Z., Xia, L., Yu, Y., Ao, T., Huang, C.: LightRAG: simple and fast retrieval-augmented generation. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2025, pp. 10746–10761. Association for Computational Linguistics, Suzhou, China (2025). https://doi.org/10.18653/v1/2025.findings-emnlp.568. https://aclanthology.org/2025.findings-emnlp.568/

6. Gutiérrez, B.J., Shu, Y., Qi, W., Zhou, S., Su, Y.: From RAG to memory: non-parametric continual learning for large language models. In: Forty-second International Conference on Machine Learning (2025). https://openreview.net/forum?id=LWH8yn4HS2
7. Ha, N.T., et al.: Vietnamese legal question answering: an experimental study. In: 2024 16th International Conference on Knowledge and System Engineering (KSE), pp. 440–446 (2024). https://doi.org/10.1109/KSE63888.2024.11063637
8. Lee, J., Kim, D., Hwang, S., Kim, H., Lee, G.: KoBLEX: open legal question answering with multi-hop reasoning. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pp. 4019–4053. Association for Computational Linguistics, Suzhou, China (2025). https://doi.org/10.18653/v1/2025.emnlp-main.200, https://aclanthology.org/2025.emnlp-main.200/
9. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)
10. Li, D., et al.: From generation to judgment: opportunities and challenges of LLM-as-a-judge. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pp. 2757–2791. Association for Computational Linguistics, Suzhou, China (2025). https://doi.org/10.18653/v1/2025.emnlp-main.138, https://aclanthology.org/2025.emnlp-main.138/
11. de Martim, H.: Graph RAG for legal norms: a hierarchical and temporal approach (2025)
12. Martinez-Gil, J.: A survey on legal question–answering systems. Comput. Sci. Rev. **48**, 100552 (2023). https://doi.org/10.1016/j.cosrev.2023.100552, https://www.sciencedirect.com/science/article/pii/S1574013723000199
13. Minh, D.D., Van, V.N., Cong, T.D.: Using large language models for education managements in vietnamese with low resources. In: Oco, N., Dita, S.N., Borlongan, A.M., Kim, J.B. (eds.) Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation, pp. 20–34. Tokyo University of Foreign Studies, Tokyo, Japan (2024). https://aclanthology.org/2024.paclic-1.3/
14. Nguyen, L.S.T., et al.: Can small language models handle vietnamese legal reasoning? insights from multi-task evaluation. In: Mai, L.C., Huyen, N.T.M., Trang, N.T.T. (eds.) Proceedings of the 11th International Workshop on Vietnamese Language and Speech Processing, pp. 165–178. Association for Computational Linguistics, Hanoi, Vietnam (2025). https://aclanthology.org/2025.vlsp-1.23/
15. Nguyen, N.T.H., Ha, P.P.D., Nguyen, L.T., Van Nguyen, K., Nguyen, N.L.T.: SP-BERTQA: a two-stage question answering system based on sentence transformers for medical texts. In: Knowledge Science, Engineering and Management, pp. 371–382. Springer International Publishing (2022)
16. Peng, B., et al.: Graph retrieval-augmented generation: a survey. ACM Trans. Inf. Syst. (2025). https://doi.org/10.1145/3777378, just Accepted
17. Pham, V., Le, H.H., Ngo, T.P., Nguyen, B., Nguyen, D., Nguyen, H.D.: Enhancing legal research through knowledge-infused information retrieval for vietnamese labor law. IAES Int. J. Artif. Intell. (IJ-AI) **13**(4), 3962–3973 (2024). https://doi.org/10.11591/ijai.v13.i4.pp3962-3973, https://ijai.iaescore.com/index.php/IJAI/article/view/25255

18. Sarthi, P., Abdullah, S., Tuli, A., Khanna, S., Goldie, A., Manning, C.D.: RAPTOR: recursive abstractive processing for tree-organized retrieval. In: The Twelfth International Conference on Learning Representations (2024). https://openreview.net/forum?id=GN921JHCRw
19. Tran, M.N., Nguyen, P.V., Nguyen, L., Dinh, D.: ViMedAQA: a vietnamese medical abstractive question-answering dataset and findings of large language model. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, pp. 252–260. Association for Computational Linguistics (2024). https://doi.org/10.18653/v1/2024.acl-srw.31. https://aclanthology.org/2024.acl-srw.31/
20. Trivedi, H., Balasubramanian, N., Khot, T., Sabharwal, A.: interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 10014–10037. Association for Computational Linguistics, Toronto, Canada (2023). https://doi.org/10.18653/v1/2023.acl-long.557. https://aclanthology.org/2023.acl-long.557/
21. Van Nguyen, K., Van Huynh, T., Nguyen, D.V., Nguyen, A.G.T., Nguyen, N.L.T.: New vietnamese corpus for machine reading comprehension of health news articles. ACM Trans. Asian Low-Res. Lang. Inf. Process. **21**(5) (2022).https://doi.org/10.1145/3527631