information

MDPI

Article

An Empirical Study of Knowledge Graph-Enhanced RAG for Information Security Compliance

Dimitar Jovanovski *, Marija Stojcheva, Mila Dodevska, Petre Lameski, Igor Mishkovski and Dejan Gjorgjevikj

Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, 1000 Skopje, North Macedonia; marija.stojcheva@finki.ukim.mk (M.S.); mila.dodevska@finki.ukim.mk (M.D.); petre.lameski@finki.ukim.mk (P.L.); igor.mishkovski@finki.ukim.mk (I.M.); dejan.gjorgjevikj@finki.ukim.mk (D.G.)

* Correspondence: dimitar.jovanovski.1@students.finki.ukim.mk

Abstract

Information security compliance has become critical for organizations worldwide, with the ISO/IEC 27000 family serving as the most widely adopted framework for establishing information security management systems. Despite their global acceptance, these standards present significant interpretation challenges due to their formal language, abstract structure, and extensive cross-referencing across 97 documents. Traditional retrieval-augmented generation (RAG) systems, which rely on independent text chunking and dense vector retrieval, prove inadequate for such highly interconnected regulatory materials, often fragmenting contextual relationships and reducing accuracy. This study introduces a privacy-preserving RAG framework that integrates LightRAG, a knowledge graph-based retrieval system, with locally hosted open-source language models. Unlike chunk-based RAG systems that treat document segments independently, the system in this study constructs a semantic knowledge graph that explicitly models relationships between clauses through typed edges representing cross-references, semantic similarity, and hierarchical dependencies. To enable rigorous evaluation, we developed a curated benchmark dataset of 222 multiple-choice questions with authoritative ground-truth answers, systematically constructed from official ISO standards, certification preparation materials, and academic sources. Through systematic evaluation on this benchmark, we show that knowledge graph-based retrieval achieves higher accuracy than chunk-based RAG and non-retrieval LLM baselines within the evaluated setup. The analysis indicates that embedding model quality is strongly associated with system performance, that hybrid retrieval modes combining local and global graph traversal tend to yield better accuracy, and that mid-sized open-source models paired with strong retrievers can approach the performance of larger proprietary systems. The best configuration achieves 90.54% accuracy, demonstrating the promising effectiveness of graph-structured retrieval for multiple-choice regulatory questions.

Keywords: retrieval-augmented generation; knowledge graph; information security; ISO/IEC 27000 standards; compliance; open-source models

Academic Editors: Marina Bagic Babac and Andrea Giovanni Nuzzolese

Received: 14 February 2026
Revised: 4 April 2026
Accepted: 5 April 2026
Published: 20 April 2026

1. Introduction

Today, information security is a core component of digital risk management, as organizations increasingly depend on digital infrastructures to create, store, and process sensitive data. The scale of cyber threats, regulatory obligations, and data protection

requirements has elevated information security from a purely technical concern to an organizational and governance challenge. Among the standards used to structure information security practice, the ISO/IEC 27000 series is one of the most widely known and adopted frameworks for information security management. In this study, however, the evaluation is limited to a subset of seven widely used standards from that series. These standards cover terminology, management requirements, security controls, implementation guidance, measurement approaches, and certification criteria. Together, these standards provide a structured foundation for establishing and maintaining an information security management system.

While the ISO/IEC 27000 standards provide comprehensive guidance, they also bring substantial practical challenges. The documents are highly formal, conceptually abstract, and densely interconnected, often cross-referencing clauses, sections, and companion standards in order to interpret a single requirement correctly. As a result, practitioners, auditors, and engineers may struggle to identify the relevant provisions and to integrate related evidence spread across multiple parts of the corpus. Misinterpretation or incomplete interpretation can lead to compliance gaps, inconsistent implementation, and increased organizational risk.

Retrieval-augmented generation (RAG) has emerged as a promising approach for assisting with the interpretation of complex regulatory and technical documentation. However, conventional RAG systems typically rely on fixed-size text chunking and dense vector similarity, which are not compatible with the structure of regulatory texts. In documents such as the ISO/IEC 27000 standards, relevant evidence is often distributed across hierarchically related sections, explicit clause cross-references, and semantically linked concepts spanning multiple documents. Chunk-based retrieval fragments context and overlooks inter-clause relationships that are necessary for correct interpretation. In contrast, knowledge graph-enhanced retrieval is a plausible alternative for regulatory question answering because it can preserve structural and semantic relationships between clauses, concepts, and documents, thereby supporting retrieval that may be better suited to cross-referential and multi-hop evidence gathering.

To address these challenges, this study makes four contributions. First, we formulate multiple-choice question answering over a subset of seven ISO/IEC 27000-series standards as a knowledge graph-enhanced RAG task, motivated by the need to preserve hierarchical, cross-referential, and semantic relationships rather than relying solely on isolated text chunks. Second, we implement a privacy-preserving local architecture based entirely on open-source components, illustrating how such a pipeline can be constructed without relying on external cloud services. Third, we introduce a curated benchmark of 222 multiple-choice questions with authoritative ground-truth answers derived from official ISO standards, certification preparation materials, and academic sources, providing a reproducible evaluation resource for compliance-oriented AI research. Fourth, through a systematic benchmark-based evaluation across embedding models, reasoning models, retrieval modes, and system parameters, we examine how retrieval configuration is associated with downstream answer accuracy within the evaluated setup.

The experimental study is guided by three research questions: (1) Within this seven-standard benchmark, does knowledge graph-enhanced retrieval achieve higher multiple-choice answer accuracy than the naïve retrieval baseline used in this study? (2) How are benchmark accuracy results affected by the choice of embedding model, reasoning model, and retrieval mode within this domain? (3) How do system parameters such as chunk size and context window affect multiple-choice answer accuracy in the evaluated RAG pipeline?

The remainder of this paper is structured as follows. Section 2 reviews related work on RAG systems, knowledge graphs, and compliance automation. Section 3 describes the

proposed methodology, including the system architecture and benchmark construction. Section 4 presents the experimental results. Section 5 discusses the findings, limitations, and implications. Section 6 concludes the paper and outlines directions for future work.

The scope of the study is intentionally limited to benchmark-based multiple-choice question answering over seven ISO/IEC 27000-series standards, and the findings should not be interpreted as direct evidence of performance on the broader standard family, sector-specific extensions, or real-world compliance practice.

2. Related Work

Existing research on AI-assisted compliance has largely focused on leveraging language models and retrieval systems to assist in the interpretation of regulatory and technical documents. Several studies demonstrate that RAG architectures can reduce hallucinations and improve factual grounding when applied to complex, domain-specific texts [1–3] The foundational RAG architecture, introduced by Lewis et al. [4], combines dense retrieval with neural generation to improve the factual accuracy in knowledge-intensive tasks. Subsequent work has extended this paradigm through self-reflection mechanisms, query rewriting, adaptive retrieval strategies, and integration with extremely long context windows [5–8]. Comprehensive surveys by Gao et al. [9] and Fan et al. [10] provide extensive taxonomies of RAG architectures and identify retrieval quality as a critical bottleneck. Such systems have shown promising results in domains such as financial services, data protection, medical compliance, and policy management, where regulatory documents are lengthy, cross-referential, and difficult to interpret without automated support [11–13].

Despite these advances, most prior solutions still rely on conventional chunk-based retrieval, which splits the documents into fixed-size chunks that are treated as independent units and stored in vector databases, and retrieved based on embedding similarity. While effective for well-structured or narrative text, this method often proves insufficient for documents with strong internal cross-references and hierarchical dependencies, such as the ISO/IEC standards. Because each chunk is indexed independently, important semantic relationships across clauses are lost, leading to fragmented context and reduced retrieval precision. This limitation has been noted in multiple studies, which argue that compliance-driven RAG applications require explicit modeling of relationships rather than relying solely on similarity-based text matching [14–17].

Further studies have investigated locally deployed and privacy-preserving RAG systems. Given that compliance documents often include sensitive organizational information, cloud-based RAG solutions introduce concerns related to data confidentiality. In response, researchers have proposed federated, on-device, and fully local RAG architectures designed to preserve data sovereignty while enabling regulatory analysis [18–23]. Complementary work on federated learning [24,25] and differential privacy [26] has demonstrated techniques for collaborative model training without centralizing sensitive data, although the direct application to RAG systems remains an open challenge.

The RAG system proposed in this study builds upon these limitations by incorporating knowledge graph-based retrieval. Integration of knowledge graphs with large language models has emerged as a promising research direction, with recent work demonstrating that explicit structural representations can improve reasoning, reduce hallucination, and enable more interpretable output [27,28]. Microsoft's GraphRAG system [29] demonstrated that graph-based retrieval can significantly improve query-focused summarization by capturing both local and global document structure. Unlike traditional chunk-based systems, LightRAG constructs a semantic graph in which nodes represent clauses or concepts and edges capture relationships such as cross-references, thematic similarity, or hierarchical structure [30]. This design enables retrieval that reflects both lexical proximity

and structural relationships, allowing the system to navigate the deeper interconnectedness of regulatory materials. As a result, the proposed approach produces context that more accurately captures how ISO/IEC 27000 clauses relate to each other, thereby improving both recall and interpretability.

3. Methodology

This section outlines the methodological approach used to evaluate a locally deployable, knowledge graph-enhanced retrieval-augmented generation (RAG) system for question answering over the ISO/IEC 27000 family of standards. The methodology is structured in two parts. The first describes the system architecture, including the integration of LightRAG with locally hosted Ollama models and the retrieval-generation pipeline. The second describes the preparation of the ISO corpus and the construction of the benchmark dataset used for evaluation. Together, these methodological components are designed to support the study's research questions concerning the contribution of graph-enhanced retrieval, model selection, and system parameterization to regulatory question answering performance.

3.1. System Design and System Architecture

The goal of the proposed system is to enable accurate, privacy-preserving reasoning over the ISO/IEC 27000 family of standards by combining knowledge graph-enhanced retrieval with locally hosted open-source language models. The architecture is built around LightRAG [30], which extends conventional vector retrieval with a graph representation derived from the document corpus. Rather than treating each text chunk as an isolated retrieval unit, the system first segments the documents into clause- or paragraph-level chunks and then uses a language model to extract entities, concepts, and typed relationships from those chunks. These extracted elements form a corpus-level knowledge graph enriched with metadata linking each node and relation back to its source segment. In the context of the ISO/IEC 27000 standards, this representation is intended to preserve document hierarchy, explicit cross-references, and semantically related concepts that may be distributed across multiple clauses or documents. This is done in three main stages:

1. Document ingestion and graph construction: The source documents are first segmented into clause or paragraph-level chunks. Each chunk is embedded and stored in a vector index for semantic retrieval. In parallel, LightRAG prompts a language model to extract entities and typed relationships from each chunk, which are then assembled into a corpus-level knowledge graph. The extracted relations are stored together with metadata that links them back to the originating chunk(s). Repeated entities and relations extracted from different parts of the corpus are deduplicated, thereby linking conceptually related content across documents. Importantly, graph edges are not created using an embedding-similarity threshold; instead, cosine similarity is used only during retrieval from the vector index to identify candidate chunks, entities, or relations. An example of a knowledge graph constructed in this manner is shown in Figure 1.
2. Retrieval by querying the knowledge graph: When a user submits a prompt, LightRAG can operate in several retrieval modes. In naïve mode, retrieval is based only on vector similarity over chunk embeddings. In local mode, retrieval focuses on graph elements closely associated with the entities and relations most directly relevant to the query, aiming to recover precise clause-level evidence. In global mode, retrieval expands over broader graph neighborhoods to collect more conceptually distributed context. In hybrid mode, the system combines both local and global evidence. This retrieval design is particularly relevant for regulatory corpora, where some questions

depend on a specific clause, while others require combining information across linked paragraphs, sections, and/or standards.

3. Grounded answer generation: The retrieved evidence, together with the user query, is provided to a locally hosted Ollama language model, which generates the final answer. Because answer generation is conditioned on retrieved ISO-related context rather than on the model's parametric memory alone, the resulting response remains grounded and factual.

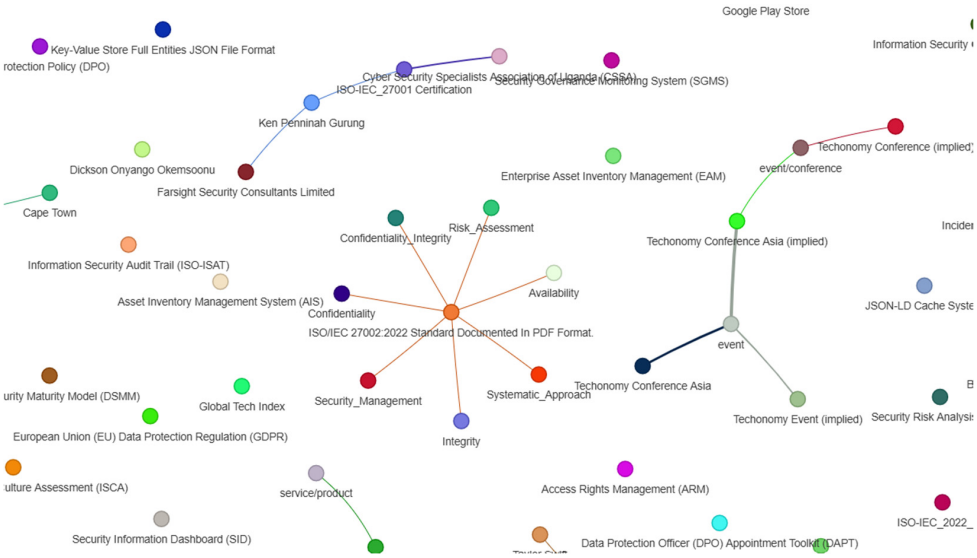

Figure 1. Sample of the constructed knowledge graph from the subset of ISO 27000 documents in the study.

Because all components of this system, including the language and embedding models, run fully locally, the pipeline ensures data confidentiality and full reproducibility. The LLM synthesizes a response explicitly grounded in retrieved ISO content, reducing hallucinations and improving factual consistency.

To examine how system configuration affects performance, several internal parameters were varied alongside the embedding model, reasoning model, and retrieval mode. The most important configurable parameters were:

1. NUM_CTX: The maximum number of tokens available to the reasoning model during answer generation [31];
2. EMBEDDING_DIM: The dimensionality of the vector representations produced by the selected embedding model [31];
3. MAX_EMBED_TOKENS: The maximum chunk length allowed prior to embedding, which influences retrieval granularity and context segmentation [31].

These parameters were selected because they affect the amount of context available for reasoning, the structure of the indexed representation, and the balance between localized and broader retrieval. The modular design of the system allows these variables to be varied independently, supporting systematic comparison across multiple configurations. All experiments were conducted on the same workstation equipped with an NVIDIA RTX 3090 GPU in order to maintain consistent hardware conditions across runs. A high-level overview of the full pipeline is shown in Figure 2.

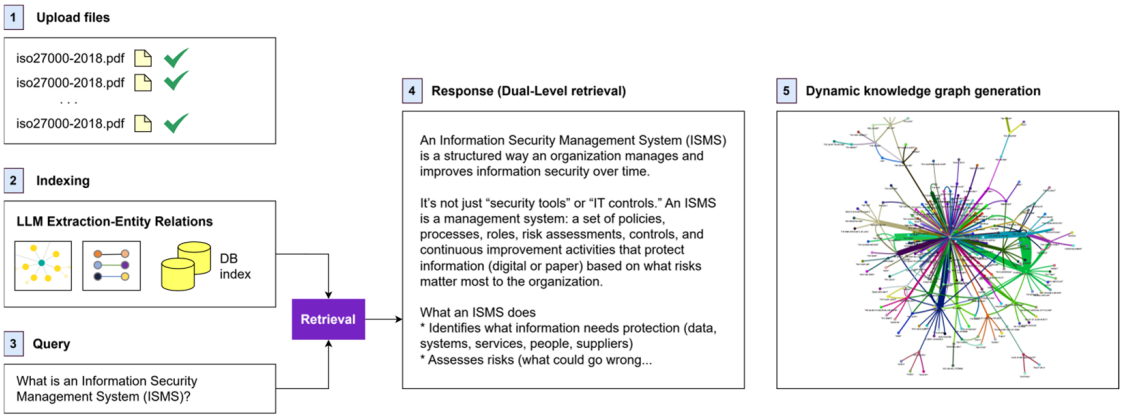

Figure 2. High-level view of LightRAG's workflow pipeline.

3.2. ISO Corpus Preparation and Benchmark Dataset

The knowledge base for this study consists of a subset of the ISO/IEC 27000 family of standards focused on information security management. It is intentionally restricted to provide broad coverage of core information security management concepts while remaining computationally manageable for systematic experimentation. The following documents are included in the study:

- ISO/IEC 27000:2018 [32]: Overview and vocabulary;
- ISO/IEC 27001:2022 [33]: Requirements for ISMS;
- ISO/IEC 27002:2022 [34]: Code of practice for information security controls;
- ISO/IEC 27003:2017 [35]: Implementation guidance for ISMS;
- ISO/IEC 27004:2016 [36]: Measurement of information security;
- ISO/IEC 27005:2022 [37]: Information security risk management;
- ISO/IEC 27006:2024 [38]: Requirements for bodies auditing and certifying ISMS.

While this study evaluates retrieval-augmented generation over the above-mentioned seven core ISO/IEC 27000 standards, the findings cannot be generalized to the full ISO/IEC 27000 family of 97 documents or to sector-specific extensions such as ISO/IEC 27701, 27017, or 27018. The results should therefore be interpreted strictly within the context of this selected subset and the multiple-choice benchmark used in the experiments.

In our case, we convert each PDF into structured machine-readable text. Preprocessing included removing headers and footers, normalizing numbering formats, reconstructing broken paragraphs, and ensuring consistent clause and section labeling. The cleaned documents are then segmented into individual meaning units, typically at the clause or paragraph level. Metadata such as clause identifiers, document names, and section headers is stored for each segment to support fine-grained traceability between retrieved evidence and the original source material.

To evaluate the system's performance, we construct a 222 multiple-choice question benchmark dataset. The questions are curated from four source categories: (1) the ISO documents listed above, (2) certification exam preparation books, (3) academic sources, and (4) industry training materials focused on the ISO/IEC 27000 standards [39–54]. Candidate questions are reviewed manually, reformulated where necessary for consistency, and paired with a single ground-truth answer derived from authoritative ISO text. The benchmark covers a broad range of topics, including terminology, security controls, risk management principles, ISMS operation, audit requirements, performance evaluation, and implementation guidance. While the dataset is intended to achieve broad topical coverage, it was not designed as a perfectly balanced psychometric instrument; rather, its purpose is to provide a reproducible domain-specific benchmark for comparative system evaluation.

A sample of the benchmark is shown in Table 1. The full dataset has been publicly released to support reproducibility and future comparison at the link: https://huggingface.co/datasets/dimitarjovanovski/ISO27K-QnA-Benchmark-dataset (accessed on 3 April 2026).

Table 1. Sample of the custom ISO/IEC 27000 family of standards evaluation dataset.

Question	Answer_gt
What is an Information Security Management System (ISMS)? (A) A set of policies and procedures for managing sensitive company information. (B) A software tool for managing security risks (C) A physical security system (D) A consulting service for security compliance	A
What are the benefits of implementing ISO 27001? (A) Improving an organization's overall security posture (B) Enhancing an organization's reputation and credibility (C) Facilitating compliance with legal and regulatory requirements (D) All of the above	D
What is the main difference between ISO 27001 and ISO 27002? (A) ISO 27001 is a standard and ISO 27002 is a code of practice (B) ISO 27001 is for management and ISO 27002 is for technical implementation (C) ISO 27001 is for small businesses and ISO 27002 is for large organizations (D) ISO 27001 is for government agencies and ISO 27002 is for the private sector	A

The benchmark was designed to support consistent automated comparison across many retrieval and model configurations, rather than to serve as a comprehensive evaluation of real-world compliance reasoning. Accordingly, it does not assess free-form answer generation, reasoning-trace quality, citation faithfulness, or robustness to paraphrase and distractor variation. While candidate questions were manually reviewed and aligned to a single reference answer grounded in ISO text, the current version of the benchmark was not accompanied by formal inter-annotator agreement analysis or expert-user validation. In addition, the use of multiple-choice questions simplifies the reasoning task and may overstate accuracy compared to open-ended QA formats. The inclusion of questions from secondary sources, such as exam preparation books and industry training materials, may introduce systematic bias. These limitations mean that the reported accuracy should be interpreted as exploratory rather than definitive evidence of compliance reasoning performance, and future work should incorporate larger datasets, expert validation, and evaluation of reasoning-trace quality and robustness. Therefore, the findings should be interpreted strictly as evidence of effectiveness in multiple-choice regulatory QA, not as validation of performance in complex regulatory interpretation or real compliance practice.

Together, the ISO corpus and the multiple-choice benchmark enable systematic, quantitative assessment of the retrieval and reasoning capabilities of the knowledge graph-based RAG system under varying embedding models, language models, retrieval modes, and parameter configurations.

The multiple-choice format was selected to enable consistent automated scoring across a large number of configurations, although this format simplifies the task for the language model relative to open-ended generation. Scoring and validation based on full-text, open-ended responses are left for future work, as they require more complex evaluation procedures and potentially expert-assisted assessment.

The experimental design varies retrieval-related settings and reasoning models separately. Retrieval is varied through the choice of embedding model, retrieval mode, and chunking parameters, while generation is varied through the choice of reasoning model.

This design supports comparative analysis of their association with final-answer accuracy, but it does not fully disentangle retrieval effects from generation effects.

4. Results

This section presents the experimental results in relation to the study's three research questions. In the first phase, we examine the effect of the embedding model, reasoning model, and retrieval mode on answer accuracy. And in the second phase, we evaluate how system parameters, specifically chunk size and context window, influence performance for the best-performing retrieval configuration.

The results below cover the research questions from the previous sections by showing that graph-enhanced retrieval generally improves performance over naïve retrieval and that the embedding model contributes more strongly to performance variation than the downstream reasoning model.

4.1. First Phase: Embedding and Model Selection

In the first phase, the accuracy of every embedding–reasoning model combination was measured three times, averaged out, and ranked. As shown in Table 2, the embedding model and retrieval mode appear to have a stronger influence on overall performance than the choice of reasoning model within the evaluated configurations. This observation aligns with findings from Izacard et al. (2022) [55], which demonstrate that smaller language models paired with strong retrievers can match or exceed the performance of much larger language models lacking retrieval support.

In our experiments, the mxbai-embed-large:335m embedding model [56] achieved the highest overall accuracy. The nomic-embed-text:137m model [57] performed similarly and ranked second. By contrast, the embeddinggemma:300m model [58] produced significantly lower accuracy, despite having nearly twice as many parameters as nomic-embed-text:137m. These results further highlight that embedding quality, rather than embedding model size, is the critical factor in retrieval-centric architectures.

Among the stronger embedding configurations shown in Table 2, hybrid retrieval, which combines both local entity-focused and global concept-level traversal, is usually the best-performing or tied-best retrieval mode, while global retrieval is also a strong competitor in several cases. In contrast, with the weakest embedding model (embeddinggemma:300m), the benefits of graph-enhanced retrieval are less stable, suggesting that retrieval strategy and embedding quality interact rather than contributing independently. This finding demonstrates that leveraging both fine-grained relationships and broader conceptual connections enables more comprehensive context retrieval, particularly for questions requiring integration of information from multiple document sections or standards.

Table 2 reveals three main patterns. First, mxbai-embed-large:335m produces the strongest overall results, with peak accuracies of 83.78%. Second, nomic-embed-text:137m shows stable but slightly lower performance, peaking around–79.73–80.18% depending on retrieval mode. Third, embeddinggemma:300m produces substantially more variable results, indicating that retrieval mode alone cannot compensate for weaker embeddings.

Since each question presents four answer options, a random baseline would achieve an expected accuracy of 25%. All evaluated configurations perform substantially above this chance level.

Table 2. Averaged accuracy results of three runs with their standard deviations and confidence intervals for different embedding–reasoning model combinations and retrieval modes.

Embedding Model	Language Model	Retrieval modes			
		Naïve	Local	Hybrid	Global
nomic-embed-text:137m	deepseek-r1:8b	78.38% ± 0.45% [77.26, 79.50]	78.83% ± 0.90% [76.59, 81.07]	80.18% ± 0.78% [78.24, 82.12]	79.28% ± 0.90% [77.04, 81.52]
	llama3.1:8b	78.38% ± 0.45% [77.26, 79.50]	78.38% ± 0.45% [77.26, 79.50]	80.18% ± 0.78% [78.24, 82.12]	79.28% ± 0.90% [77.04, 81.52]
	mistral-nemo:12b	76.58% ± 0.78% [74.64, 78.52]	78.38% ± 0.45% [77.26, 79.50]	80.18% ± 0.78% [78.24, 82.12]	79.28% ± 0.90% [77.04, 81.52]
	qwen2.5:14b	77.48% ± 0.45% [76.36, 78.60]	78.38% ± 0.45% [77.26, 79.50]	80.18% ± 0.78% [78.24, 82.12]	79.73% ± 1.19% [76.77, 82.69]
	gpt-oss:20b	77.48% ± 0.45% [76.36, 78.60]	78.38% ± 0.45% [77.26, 79.50]	80.18% ± 0.78% [78.24, 82.12]	79.73% ± 1.19% [76.77, 82.69]
embeddinggemma:300m	deepseek-r1:8b	61.71% ± 0.90% [59.47, 63.95]	55.41% ± 1.35% [52.06, 58.76]	53.15% ± 1.80% [48.68, 57.62]	51.80% ± 2.25% [46.21, 57.39]
	llama3.1:8b	68.47% ± 0.90% [66.23, 70.71]	59.91% ± 1.35% [56.56, 63.26]	56.76% ± 1.81% [52.26, 61.26]	71.62% ± 2.25% [66.03, 77.21]
	mistral-nemo:12b	68.92% ± 0.45% [67.80, 70.04]	64.86% ± 0.90% [62.62, 67.10]	62.16% ± 1.35% [58.81, 65.51]	72.07% ± 1.80% [67.60, 76.54]
	qwen2.5:14b	66.22% ± 0.45% [65.10, 67.34]	74.32% ± 0.90% [72.08, 76.56]	65.32% ± 0.91% [63.06, 67.58]	72.07% ± 1.80% [67.60, 76.54]
	gpt-oss:20b	61.26% ± 0.45% [60.14, 62.38]	61.26% ± 0.90% [59.02, 63.50]	66.22% ± 0.90% [63.98, 68.46]	71.62% ± 1.80% [67.15, 76.09]
mxbai-embed-large:335m	deepseek-r1:8b	79.73% ± 0.45% [78.61, 80.85]	78.83% ± 2.25% [73.24, 84.42]	81.53% ± 0.90% [79.29, 83.77]	81.08% ± 1.35% [77.73, 84.43]
	llama3.1:8b	81.08% ± 0.00% [81.08, 81.08]	78.38% ± 0.45% [77.26, 79.50]	81.53% ± 0.90% [79.29, 83.77]	80.63% ± 1.35% [77.28, 83.98]
	mistral-nemo:12b	81.08% ± 0.00% [81.08, 81.08]	79.27% ± 0.45% [78.15, 80.39]	82.88% ± 0.90% [80.64, 85.12]	81.98% ± 1.19% [79.02, 84.94]
	qwen2.5:14b	81.08% ± 0.45% [79.96, 82.20]	78.83% ± 0.45% [77.71, 79.95]	83.33% ± 0.90% [81.09, 85.57]	83.33% ± 1.35% [79.98, 86.68]
	gpt-oss:20b	81.08% ± 0.90% [78.84, 83.32]	78.83% ± 0.90% [76.59, 81.07]	83.33% ± 0.45% [82.21, 84.45]	83.78% ± 0.90% [81.54, 86.02]

4.2. Second Phase: Parameter Optimization

In the second phase, the mxbai-embed-large:335m model was selected as the best embedding model based on its superior performance in Phase 1. Three language models were chosen for further evaluation: deepseek-r1:8b [59], qwen2.5:14b [60], and gpt-oss:20b [61]. All models in this phase used the hybrid retrieval mode, as it combines both local and global graph traversal and consistently produced the highest accuracy in Phase 1.

To investigate the impact of system parameters, different values of NUM_CTX (the model's context window size) and MAX_EMBED_TOKENS (the maximum token length per embedded chunk) were tested. The results in Table 3 show a clear overall trend: smaller or medium-sized embedding chunks combined with larger context windows generally outperform larger chunk sizes. The best overall configuration was MAX_EMBED_TOKENS = 1024 and NUM_CTX = 8192 for gpt-oss:20b, while qwen2.5:14b achieved its highest score with MAX_EMBED_TOKENS = 2048 and NUM_CTX = 8192.

This performance pattern indicates that finer-grained embeddings preserve semantic precision during retrieval, while larger context windows enable the language model to effectively process the retrieved information. In contrast, excessively large embedding chunks (MAX_EMBED_TOKENS = 8192) degraded performance across all models, since semantically distinct concepts within a single chunk reduce retrieval precision.

Table 3. Averaged accuracy, standard deviation and confidence intervals of configurations using mxbai-embed-large:335m with hybrid retrieval and different LLMs, context window sizes, and embedding token limits.

Language Model	NUM_CTX	MAX_EMBEDDING_TOKENS			
		1024	2048	4096	8192
deepseek-r1:8b	4096	83.78% ± 0.90% [81.54, 86.02]	82.88% ± 0.90% [80.64, 85.12]	81.98% ± 1.35% [78.63, 85.33]	80.63% ± 1.80% [76.16, 85.10]
	8192	86.04% ± 0.45% [84.92, 87.16]	84.23% ± 0.90% [81.99, 86.47]	83.78% ± 0.90% [81.54, 86.02]	81.53% ± 1.35% [78.18, 84.88]
qwen2.5:14b	4096	85.14% ± 0.45% [84.02, 86.26]	86.94% ± 0.45% [85.82, 88.06]	85.14% ± 0.90% [82.90, 87.38]	82.43% ± 1.35% [79.08, 85.78]
	8192	87.39% ± 0.00% [87.39, 87.39]	88.29% ± 0.45% [87.17, 89.41]	86.49% ± 0.45% [85.37, 87.61]	83.33% ± 0.90% [81.09, 85.57]
gpt-oss:20b	4096	88.29% ± 0.45% [87.17, 89.41]	82.88% ± 0.90% [80.64, 85.12]	83.33% ± 0.90% [81.09, 85.57]	82.43% ± 1.35% [79.08, 85.78]
	8192	90.54% ± 0.00% [90.54, 90.54]	88.74% ± 0.45% [87.62, 89.86]	86.04% ± 0.45% [84.92, 87.16]	83.33% ± 0.90% [81.09, 85.57]

In order to cover all cases, as a final step, we analyzed and compared the performances of the three best LightRAG configurations presented in Table 3 with the following large language models (without an external knowledge base):

- deepseek-r1:8b
- qwen2.5:14b
- gpt-oss:20b
- gpt-5.2

The accuracy results achieved by the large language models are the average of three measurements, in order to ensure reliability and eliminate evaluation bias.

As shown in Figure 3, the standalone language models achieved substantially lower accuracy than the top LightRAG configurations, confirming that access to external structured evidence is more important than parametric model scale alone in this benchmark. This confirms once more that the use of an external knowledge base and the quality of retrieval appear to have a greater impact on the final performance of the system than the size of the language model (its number of parameters).

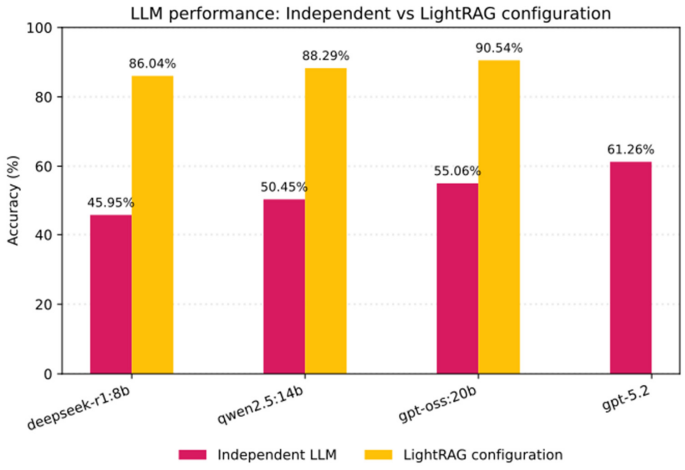

Figure 3. Performance comparison between the top three LightRAG configurations and standalone LLMs.

5. Discussion

Within the evaluated benchmark, graph-enhanced retrieval generally achieved higher mean multiple-choice answer accuracy than the naïve similarity-based retrieval baseline used in this study. Across the strongest evaluated model combinations, the graph-enhanced configurations obtained accuracies between 81.53% and 90.54%, compared with 78.38% to 81.08% for the naïve retrieval baseline, yielding observed differences of approximately 3–10 percentage points in the best comparisons.

Retrieval quality, driven by the embedding model and retrieval strategy, appears to have a greater impact on system performance than the size or complexity of the downstream reasoning model within the evaluated configurations. Among the tested embeddings, mxbai-embed-large:335m yields the strongest overall performance, surpassing both nomic-embed-text:137m and embeddinggemma:300m, despite having fewer parameters than the latter. This result suggests that, for compliance-oriented question answering, investing in higher-quality embedding models and retrieval infrastructure is more important than the standalone language model's size.

The results further suggest that retrieval strategies combining vector-based search with graph-based context expansion are beneficial for regulatory question answering. In the stronger embedding configurations, hybrid and global retrieval modes generally outperform naïve retrieval and often outperform strictly local retrieval. This pattern is consistent with the expectation that graph-enhanced retrieval is especially useful when relevant evidence is distributed across multiple clauses, sections, or companion standards. For example, questions that require linking security controls described in ISO/IEC 27002 with risk assessment procedures in ISO/IEC 27005 are difficult to answer using isolated chunk retrieval alone, because the relevant evidence may be distributed across documents. In contrast, graph-enhanced retrieval can better preserve and exploit structural relationships between related concepts and provisions, allowing the system to return a more complete supporting context. Although graph construction and maintenance introduce additional memory and computational overhead, this trade-off is acceptable in regulatory and compliance settings where accuracy and interpretability are prioritized.

The parameter sensitivity analysis indicates that the strongest overall configurations use small or medium embedding chunks together with large context windows, with the best overall result obtained at MAX_EMBED_TOKENS = 1024 and NUM_CTX = 8192. This pattern reflects a trade-off between semantic precision and synthesis capacity. Smaller embedding chunks preserve semantic specificity by ensuring that each embedded unit corresponds more closely to a coherent concept, thereby reducing retrieval noise and false positives. Finer-grained chunking, however, increases context fragmentation and therefore benefits from larger context windows that support integration across retrieved passages. In contrast, larger embedding chunks dilute semantic granularity, while smaller context windows constrain the model's ability to synthesize complex, multi-clause information.

Although gpt-oss:20b achieved the highest absolute accuracy in the best-performing configuration, mid-sized open-source models such as qwen2.5:14b also performed strongly under favorable retrieval settings (88.29% vs. 90.54%). Within this benchmark, this narrower gap suggests that retrieval configuration may matter at least as much as model scale for final-answer accuracy. However, the study does not evaluate latency, cost, or practitioner utility, so these results should not be interpreted as a full deployment recommendation.

However, several limitations remain. The best-performing configuration achieves 90.54% accuracy, leaving 9.46% of questions answered incorrectly, indicating that aggregate accuracy alone does not fully capture system behavior. The use of multiple-choice questions inflates measured performance relative to open-ended compliance tasks (drafting ISO/IEC-aligned justifications, assessing regulatory scenarios, or producing audit

evidence), where answer completeness, grounding, and citation quality are essential. In addition, the study does not compare system performance with human practitioners, junior auditors, or certification candidates, so the measurement of real-world significance of the observed benchmark accuracy will remain for future studies.

The study only compares LightRAG with chunk-based RAG and a non-retrieval LLM baseline. Other retrieval strategies, such as BM25 + RAG, hybrid sparse–dense retrieval, hierarchical summarization, or structured citation prompting, were not included. This leaves room for future work, which will include extended comparisons to a broader set of baselines.

However, several limitations remain. Although the benchmark results are informative at the aggregate level, accuracy alone does not fully characterize residual system behavior. To provide a more transparent exploratory analysis, we manually reviewed 10 incorrect predictions from the gpt-oss:20b configuration and 10 incorrect predictions from the deepseek-r1:8b configuration reported in Table 3 (n = 20 total). Each incorrect case was assigned one primary exploratory error label after inspection of the benchmark item, the ground-truth answer, and the model's selected answer. The observed error labels included retrieval miss, incomplete multi-hop aggregation, distractor susceptibility, and graph construction noise. Table 4 presents an illustrative subset of incorrect predictions, while Table 5 summarizes the distribution of error categories and the most frequent gold-to-predicted confusions. This analysis highlights that many errors stem from retrieval gaps or distractor sensitivity rather than reasoning alone, underscoring the need for future work with standard retrieval metrics to more precisely attribute performance gains.

Table 4. Illustrative subset of incorrect predictions from the evaluated gpt-oss:20b and deepseek-r1:8b configurations reported in Table 3.

Question	Question Type	Generated Answer	Ground Truth
What is the purpose of the Statement of Applicability according to ISO 27001?	ISMS implementation/governance	E	A
In the event of a security incident at Company X, which of the following actions should be taken first according to ISO 27001 guidelines?	Controls/Annex A	D	B
What are the criteria to be used in an internal audit of an organization's information security management system according to ISO 27001?	Audit/measurement	A	C
According to ISO 27001, Annex A, information and assets should be managed by:	Controls/Annex A	B	A
What is the purpose of ICT readiness for business continuity?	Controls/Annex A	B	D
What is the main focus of ISO/IEC 27004:2016?	Audit/measurement	C	D
The five-stage process for risk management, as laid out in ISO 27005, begins with what step?	Risk management	A	C
Which document should record the outcomes of risk assessments per ISO/IEC 27005?	Risk management	D	C

Table 5. Distribution of error categories from the evaluated gpt-oss:20b and deepseek-r1:8b configurations reported in Table 3.

Error Category	Question Type	Generated Answer
Retrieval miss	3	4
Multi-hop aggregation error	1	2
Distractor susceptibility	4	3
Graph construction noise	2	1
Total errors (subset)	10	10

Finally, although graph-based retrieval often achieved higher benchmark accuracy than the naïve baseline in Table 2, it also introduces additional computational overhead. Knowledge graph construction and storage are non-trivial, as shown in Edge et al, and frequent corpus updates may require re-extraction and re-indexing of entities and relations, which could limit scalability for very large or rapidly evolving compliance corpora. LightRAG includes mechanisms intended to support incremental graph extension, but the efficiency of such updates was not evaluated in this study.

It is important to note that the study does not provide a quantitative assessment of graph construction time, index growth, query latency, memory consumption, or incremental update costs. As a result, the practical feasibility of deploying LightRAG in production environments remains unexamined. The current findings should therefore be interpreted as methodological rather than operational, and future work should include systematic evaluation of efficiency, scalability, and resource requirements to establish deployment relevance.

6. Conclusions

This study presents a privacy-preserving RAG framework for multiple-choice question answering over a subset of seven ISO/IEC 27000-series information security standards, combining LightRAG's knowledge graph-based retrieval with locally hosted open-source models. The system constructs a semantic knowledge graph that explicitly models relationships between regulatory clauses, enabling more accurate and interpretable retrieval than conventional chunk-based approaches while ensuring full data confidentiality through on-premise deployment.

Within the evaluated benchmark, the results suggest that embedding and retrieval configuration are strongly associated with final-answer accuracy, that hybrid retrieval often performs favorably among the tested modes, and that mid-sized open-source models can perform competitively when paired with strong retrieval settings. The best observed configuration (mxbai-embed-large:335m embeddings, hybrid retrieval, 1024-token chunks, and 8192-token context windows) achieved 90.54% accuracy on the 222-question multiple-choice benchmark. These findings indicate promise for graph-enhanced retrieval in benchmark-based regulatory question answering, but they should not be interpreted as direct validation for broader compliance practice or complex real-world regulatory interpretation tasks.

An additional contribution of this work is the curated benchmark dataset of 222 multiple-choice questions with authoritative ground-truth answers, derived from official ISO standards, certification preparation materials, and academic sources. This dataset supports reproducible comparative evaluation of compliance-oriented RAG systems within a bounded benchmark setting and has been made publicly available to support future research.

Future work should address the current limitations in three main directions. First, retrieval and graph construction could be improved through adaptive chunking strategies, stronger cross-document entity linking, and incremental graph update mechanisms for

evolving regulatory corpora. Second, domain adaptation of embedding and reasoning models, including fine-tuning on compliance-oriented corpora and the use of more advanced prompting strategies, may further improve performance on complex cross-referential questions. Third, evaluation should be extended beyond multiple-choice benchmarking to include free-form question answering, scenario-based reasoning, retrieval coverage analysis, and expert human assessment in order to better characterize real-world applicability.

Author Contributions: Conceptualization, D.J. and D.G.; Methodology, D.J. and D.G.; Software, D.J.; Validation, D.G.; Writing—original draft, D.J., M.S., M.D., P.L. and I.M.; Writing—review & editing, D.J., M.S., M.D., P.L. and I.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partially financed by the Ministry of Education and Science of the Republic of North Macedonia through the project "Utilising AI and National Large Language Models to Advance Macedonian Language Capabilities".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in ISO27K-QnA-Benchmark-dataset at https://huggingface.co/datasets/dimitarjovanovski/ISO27K-QnA-Benchmark-dataset (accessed on 3 April 2026).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. *Payment Card Industry Data Security Standard (PCI DSS)*, Version 4.0; PCI Security Standards Council: Wakefield, MA, USA, 2022. Available online: https://www.pcisecuritystandards.org/document_library (accessed on 3 April 2026).
2. *Cloud Controls Matrix (CCM)*, Version 4.0; Cloud Security Alliance: Bellevue, WA, USA, 2021. Available online: https://cloudsecurityalliance.org/research/cloud-controls-matrix/ (accessed on 3 April 2026).
3. *Framework for Improving Critical Infrastructure Cybersecurity (NIST Cybersecurity Framework 2.0)*; Version 2.0; National Institute of Standards and Technology (NIST): Gaithersburg, MD, USA; U.S. Department of Commerce: Washington, DC, USA, 2024. Available online: https://www.nist.gov/cyberframework (accessed on 3 April 2026).
4. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474.
5. Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; Hajishirzi, H. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *arXiv* **2023**, arXiv:2310.11511. [CrossRef]
6. Ma, X.; Gong, Y.; He, P.; Zhao, H.; Duan, N. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 5303–5315.
7. Jeong, S.; Baek, J.; Cho, S.; Hwang, S.J.; Park, J.C. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. *arXiv* **2024**, arXiv:2403.14403.
8. Xu, P.; Ping, W.; Wu, X.; McAfee, L.; Zhu, C.; Liu, Z.; Subramanian, S.; Bakhturina, E.; Shoeybi, M.; Catanzaro, B. Retrieval meets long context large language models. *arXiv* **2023**, arXiv:2310.03025.
9. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, H.; Wang, H. Retrieval-augmented generation for large language models: A survey. *arXiv* **2023**, arXiv:2312.10997.
10. Fan, W.; Ding, Y.; Ning, L.; Wang, S.; Li, H.; Yin, D.; Chua, T.-S.; Li, Q. A survey on RAG meeting LLMs: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 6491–6501.
11. Sun, J.; Luo, Z.; Li, Y. A compliance checking framework based on retrieval augmented generation. In *Proceedings of the 31st International Conference on Computational Linguistics*; Association for Computational Linguistics: Bangkok, Thailand, 2025; pp. 2603–2615.
12. Malali, N. The Role of Retrieval-Augmented Generation (RAG) in Financial Document Processing: Automating Compliance and Reporting. *Int. J. Manag.* **2025**, *12*, 26–46. [CrossRef]
13. Han, Y.; Ceross, A.; Bergmann, J.H.M. Standard Applicability Judgment and Crossjurisdictional Reasoning: A RAG-based Framework for Medical Device Compliance. *arXiv* **2025**, arXiv:2506.18511.

14. Mao, Q.; Zhang, Q.; Hao, H.; Han, Z.; Xu, R.; Jiang, W.; Hu, Q.; Chen, Z.; Zhou, T.; Li, B.; et al. Privacy-preserving federated embedding learning for localized retrieval-augmented generation. *arXiv* **2025**, arXiv:2504.19101.
15. Addison, P.; Nguyen, M.-T.H.; Medan, T.; Shah, J.; Manzari, M.T.; McElrone, B.; Lalwani, L.; More, A.; Sharma, S.; Roth, H.R.; et al. C-FedRAG: A confidential federated retrieval-augmented generation system. *arXiv* **2024**, arXiv:2412.13163.
16. Nandagopal, S. Securing Retrieval-Augmented Generation Pipelines: A Comprehensive Framework. *J. Comput. Sci. Technol. Stud.* **2025**, *7*, 17–29. [CrossRef]
17. Rorstrom, E. *ISO 27001 Foundation—Practice Tests: 150 Questions and Explanations Based on the ISO 27001 Foundation Exam*; Kindle Direct Publishing: Seattle, WA, USA, 2023. Available online: https://www.amazon.com/ISO-27001-Foundation-Questions-Explanations-ebook/dp/B0BT21DSVT (accessed on 3 April 2026).
18. Ansari, M.S.; Khan, M.S.A.; Revankar, S.; Varma, A.; Mokhade, A.S. Lightweight Clinical Decision Support System using QLoRA-Fine-Tuned LLMs and Retrieval-Augmented Generation. *arXiv* **2025**, arXiv:2505.03406.
19. Weerasekara, T.B.; Chandeepa, C.; Amarasuriya, O.S.; Hettiarachchi, C. Privacy-Preserving Medical Advising System on Mobile Devices: On-Device PHI Anonymization, Medical Report Retrieval, and Cloud-Based RAG. In *Proceedings of the ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies*; Association for Computing Machinery: New York, NY, USA, 2025; pp. 447–452.
20. Yu, X.; Lu, Y.; Yu, Z. LocalRQA: From generating data to locally training, testing, and deploying retrieval-augmented QA systems. *arXiv* **2024**, arXiv:2403.00982.
21. Zeng, S.; Zhang, J.; He, P.; Liu, Y.; Xing, Y.; Xu, H.; Ren, J.; Chang, Y.; Wang, S.; Yin, D.; et al. The good and the bad: Exploring privacy issues in retrieval-augmented generation (RAG). In *Findings of the Association for Computational Linguistics: ACL 2024*; Association for Computational Linguistics: Bangkok, Thailand, 2024; pp. 4505–4524.
22. He, L.; Tang, P.; Zhang, Y.; Zhou, P.; Su, S. Mitigating privacy risks in Retrieval-Augmented Generation via locally private entity perturbation. *Inf. Process. Manag.* **2025**, *62*, 104150. [CrossRef]
23. Cheng, Y.; Zhang, L.; Wang, J.; Yuan, M.; Yao, Y. RemoteRAG: A privacy-preserving LLM cloud RAG service. In *Findings of the Association for Computational Linguistics: ACL 2025*; Association for Computational Linguistics: Bangkok, Thailand, 2025; pp. 3820–3837.
24. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*; PMLR: Birmingham, UK, 2017; pp. 1273–1282.
25. Kairouz, P.; McMahan, H.B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A.N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. Advances and open problems in federated learning. *Found. Trends Mach. Learn.* **2021**, *14*, 1935–8237. [CrossRef]
26. Wei, K.; Li, J.; Ding, M.; Ma, C.; Yang, H.H.; Farokhi, F.; Jin, S.; Quek, T.Q.S.; Poor, H.V. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Trans. Inf. Forensics Secur.* **2020**, *15*, 3454–3469. [CrossRef]
27. Pan, S.; Luo, L.; Wang, Y.; Chen, C.; Wang, J.; Wu, X. Unifying large language models and knowledge graphs: A roadmap. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 3580–3599. [CrossRef]
28. Sun, J.; Xu, C.; Tang, L.; Wang, S.; Lin, C.; Gong, Y.; Ni, L.M.; Shum, H.-Y.; Guo, J. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. *arXiv* **2023**, arXiv:2307.07697.
29. DEdge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; Metropolitansky, D.; Ness, R.O.; Larson, J. From local to global: A graph RAG approach to query-focused summarization. *arXiv* **2024**, arXiv:2404.16130. [CrossRef]
30. Guo, Z.; Xia, L.; Yu, Y.; Ao, T.; Huang, C. LightRAG: Simple and fast retrieval-augmented generation. *arXiv* **2024**, arXiv:2410.05779.
31. HKUDS/LightRAG Contributors. lightrag_ollama_demo.py—Example Script from LightRAG; GitHub Repository, 2025. Available online: https://github.com/HKUDS/LightRAG/blob/main/examples/lightrag_ollama_demo.py (accessed on 3 April 2026).
32. *ISO/IEC 27000:2018*; Information Technology—Security Techniques—Information Security Management Systems—Overview and Vocabulary. ISO/IEC: Geneva, Switzerland, 2018. Available online: https://www.iso.org/standard/73906.html (accessed on 3 April 2026).
33. *ISO/IEC 27001:2022*; Information Security, Cybersecurity and Privacy Protection—Information Security Management Systems—Requirements. ISO/IEC: Geneva, Switzerland, 2022. Available online: https://www.iso.org/standard/82875.html (accessed on 3 April 2026).
34. *ISO/IEC 27002:2022*; Information Security, Cybersecurity and Privacy Protection—Information Security Controls. ISO/IEC: Geneva, Switzerland, 2022. Available online: https://www.iso.org/standard/75652.html (accessed on 3 April 2026).
35. *ISO/IEC 27003:2017*; Information Technology—Security Techniques—Information Security Management Systems—Guidance. ISO/IEC: Geneva, Switzerland, 2017. Available online: https://www.iso.org/standard/63417.html (accessed on 3 April 2026).
36. *ISO/IEC 27004:2016*; Information Technology—Security Techniques—Information Security Management—Monitoring, Measurement, Analysis and Evaluation. ISO/IEC: Geneva, Switzerland, 2016. Available online: https://www.iso.org/standard/64107.html (accessed on 3 April 2026).

37. *ISO/IEC 27005:2022*; Information Security, Cybersecurity and Privacy Protection—Information Security Risk Management. ISO/IEC: Geneva, Switzerland, 2022. Available online: https://www.iso.org/standard/80585.html (accessed on 3 April 2026).
38. *ISO/IEC 27006-1:2024*; Information Security, Cybersecurity and Privacy Protection—Requirements for Bodies Providing Audit and Certification of Information Security Management Systems—Part 1: General. ISO/IEC: Geneva, Switzerland, 2024. Available online: https://www.iso.org/standard/82908.html (accessed on 3 April 2026).
39. Rorstrom, E. ISO 27001 Lead Auditor—Study Guide: Achieving Excellence in Information Security: The Ultimate ISO 27001 Lead Auditor Preparation Handbook. 2023. Available online: https://www.amazon.com/ISO-27001-Lead-Auditor-Information/dp/B0BZFC9776 (accessed on 3 April 2026).
40. Blokdyk, G. *ISO IEC 27000 A Complete Guide—2020 Edition*; 5STARCOOKS: Toronto, ON, Canada, 2020; Available online: https://books.google.mk/books?id=OZBF0AEACAAJ (accessed on 3 April 2026).
41. Wens, C. *ISO 27001 Handbook: Implementing and Auditing an Information Security Management System in Small and Medium-sized Businesses*; Independently Published, 2019. Available online: https://www.amazon.com/ISO-27001-Handbook-Implementing-medium-sized/dp/1098547683 (accessed on 3 April 2026).
42. Calder, A. *ISO27001/ISO27002: A Pocket Guide*; IT Governance Publishing: Ely, UK, 2013. Available online: https://books.google.mk/books?id=uFObBAAAQBAJ (accessed on 3 April 2026).
43. Calder, A. *ISO 27001/ISO 27002—A Guide to Information Security Management Systems*; Walter de Gruyter GmbH: Berlin, Germany, 2023; Available online: https://books.google.mk/books?id=qyHkEAAAQBAJ (accessed on 3 April 2026).
44. Hintzbergen, J.; Hintzbergen, K. *Foundations of Information Security Based on ISO27001 and ISO27002*, 3rd ed.; Van Haren Publishing: 's-Hertogenbosch, The Netherlands, 2015. Available online: https://books.google.mk/books?id=n1gdEQAAQBAJ (accessed on 3 April 2026).
45. Calder, A. *Information Security Based on ISO 27001/ISO 27002*; Van Haren Publishing: 's-Hertogenbosch, The Netherlands, 2020; Available online: https://books.google.mk/books?id=yhJADwAAQBAJ (accessed on 3 April 2026).
46. Calder, A. *Implementing Information Security Based on ISO 27001/ISO 27002*; Van Haren Publishing: 's-Hertogenbosch, The Netherlands, 2020; Available online: https://books.google.mk/books?id=0hJADwAAQBAJ (accessed on 3 April 2026).
47. Kenyon, B. *ISO 27001 Controls: A Guide to Implementing and Auditing*, 2nd ed.; IT Governance Publishing Limited: Ely, UK, 2024. Available online: https://books.google.mk/books?id=o4Hw0AEACAAJ (accessed on 3 April 2026).
48. Kyriazoglou, J. *ISO 27001: 2022 Implementation Handbook: Approaches and Measures to Comply Better with the Requirements of ISO27001 Information Security Controls Standard*; Fylatos Publishing: Thessaloníki, Greece, 2024; Available online: https://www.amazon.com/ISO-27001-Implementation-requirements-Information-ebook/dp/B0DB6FD647 (accessed on 3 April 2026).
49. *CreateSpace Independent Publishing Platform Publisher location*; CreateSpace Independent Publishing Platform: North Charleston, SC, USA, 2023. Available online: https://www.amazon.ca/Easy-Guide-Certified-ISO-27000-Specialist/dp/1542979196 (accessed on 3 April 2026).
50. Baars, H.; Hintzbergen, J.; Hintzbergen, K. *Foundations of Information Security Based on ISO27001 and ISO27002*, 4th ed.; Van Haren Publishing: 's-Hertogenbosch, The Netherlands, 2023. Available online: https://books.google.mk/books?id=xVgdEQAAQBAJ (accessed on 3 April 2026).
51. Mirtsch, M.; Kinne, J.; Blind, K. Exploring the adoption of the International Information Security Management System Standard ISO/IEC 27001: A web mining-based analysis. *IEEE Trans. Eng. Manag.* **2021**, *68*, 87–100. [CrossRef]
52. Putra, D.S.K.; Tistiyani, S.; Sunaringtyas, S.U. The Use of ISO/IEC 27001 Family of Standards in Regulatory Requirements in Some Countries. In *2021 2nd International Conference on ICT for Rural Development (IC-ICTRuDev)*; IEEE: New York, NY, USA, 2021; pp. 1–6.
53. Nowak, G.J. Information Security Management with accordance to ISO27000 Standards: Characteristics, implementations, benefits in global supply chains. *Logistyka* **2015**, *2*, 639–654.
54. de Freitas Fernandes, A.; de Brito, F.C.S.; Periard, F.F.; Matias, G.A.V.; Gonçalves, M.S.; Baldoino Filho, R.G. The ISO 27000 Family and its Applicability in LGPD Adaptation Projects for Small and Medium-Sized Enterprises. *ICSEA* **2021**, *53*.
55. Izacard, G.; Lewis, P.; Lomeli, M.; Hosseini, L.; Petroni, F.; Schick, T.; Dwivedi-Yu, J.; Joulin, A.; Riedel, S.; Grave, E. Atlas: Few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.* **2023**, *24*, 1–43.
56. Lee, S.; Shakir, A.; Koenig, D.; Lipp, J. Open source strikes bread-new fluffy embeddings model. *mixedbread* **2024**.
57. Nussbaum, Z.; Morris, J.X.; Duderstadt, B.; Mulyar, A. Nomic Embed: Training a reproducible long context text embedder. *arXiv* **2024**, arXiv:2402.01613. [CrossRef]
58. Vera, H.S.; Dua, S.; Zhang, B.; Salz, D.; Mullins, R.; Panyam, S.R.; Smoot, S.; Naim, I.; Zou, J.; Chen, F.; et al. EmbeddingGemma: Powerful and lightweight text representations. *arXiv* **2025**, arXiv:2509.20354. [CrossRef]
59. Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv* **2025**, arXiv:2501.12948.

60. Qwen, A.Y.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. Qwen2.5 technical report. *arXiv* **2024**, arXiv:2412.15115. [CrossRef]
61. Agarwal, S.; Ahmad, L.; Ai, J.; Altman, S.; Applebaum, A.; Arbus, E.; Arora, R.K.; Bai, Y.; Baker, B.; Bao, H.; et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv* **2025**, arXiv:2508.10925.