RAGulating Compliance: A Multi-Agent Schema-Light Knowledge Graph for Regulatory Compliance QA

Hemant Sunil Jomraj[1,*,†], Bhavik Agarwal[1,*,†] and Viktoria Rojkova[1,*,†]

[1]MasterControl AI Research, MasterControl, 6350 South 3000 East, Salt Lake City, Utah, USA

Abstract

Regulatory QA demands precise, verifiable answers grounded in domain text. We present a multi-agent framework that fuses a schema light (ontology minimal) knowledge graph of subject–predicate–object (SPO) triplets with retrieval-augmented generation (RAG). Agents continuously extract, normalize, and deduplicate triplets from regulatory documents; each triplet is embedded and stored, together with linked source segments and metadata, in a unified vector index. At query time, triplet level retrieval aligns user intent with concise "who-did-what-to-whom" facts and returns both the triplets and their provenance text to an LLM for answer synthesis. In complex regulatory queries, the system improves traceability and supports subgraph visualization, while achieving higher strict-threshold section overlap and better graph connectivity versus text-only baselines.

Keywords

Regulatory Compliance, Knowledge Graph, Large Language Model, Ontology Free

1. Introduction

Regulated domains (e.g., health and life sciences) demand high precision, verifiability, and domain grounding in QA[1, 2, 3]. General LLMs, including recent model families [4, 5, 6], excel in language but risk hallucinations [7, 8, 9], especially where compliance evidence and provenance are required [10]. We propose a practical system combining: (i) schema-light triplet extraction and KG maintenance [11], (ii) a unified vector store with triplets *and* source text, and (iii) a multi-agent QA pipeline that retrieves at the triplet level and returns answers with verifiable evidence.Our contributions are based on knowledge graph methods [12, 13, 14, 15] and regulatory KG/RAG applications [16, 17, 18, 19, 20, 21].

2. Method: Schema-Light KG + Triplet-Level Retrieval

2.1. Units, extraction, and provenance

The regulatory text is segmented into atomic sections (⊠: C⟶X={x$_1$, … , x_m}, then an extraction pipeline produces SPO triplets $\Phi(\Omega(C)) = \{t_i = (s_i, p_i, o_i)\}$. The provenance is captured by $\Lambda : T \to 2^X$, mapping each t_i to one or more source sections for auditability. Open IE and related practices inform the extraction side [22, 23], with open-world learning and schema emergence supported by previous work [24, 25]. Canonicalization and entity linking address vocabulary fragmentation [26, 27], while ontology-driven precedents [28] and community KGs [29, 30] motivate minimal, reusable meta-relations.

2.2. Embedding and unified index

Each triplet t_i is rendered as text $f(t_i)$, embedded via a transformer encoder into $e_{t_i} \in \mathbb{R}^d$; we store $(e_{t_i}, t_i, \Lambda(t_i))$ in a vector index. The density retrieval choices are inspired by DPR and modern similarity search [31]. Queries Q are embedded as e_Q.

ISWC 2025 Companion Volume, November 2–6, 2025, Nara, Japan

*Corresponding authors.

†These authors contributed equally.

hjomraj@mastercontrol.com (H. S. Jomraj); bagarwal@mastercontrol.com (B. Agarwal); vrojkova@mastercontrol.com (V. Rojkova)

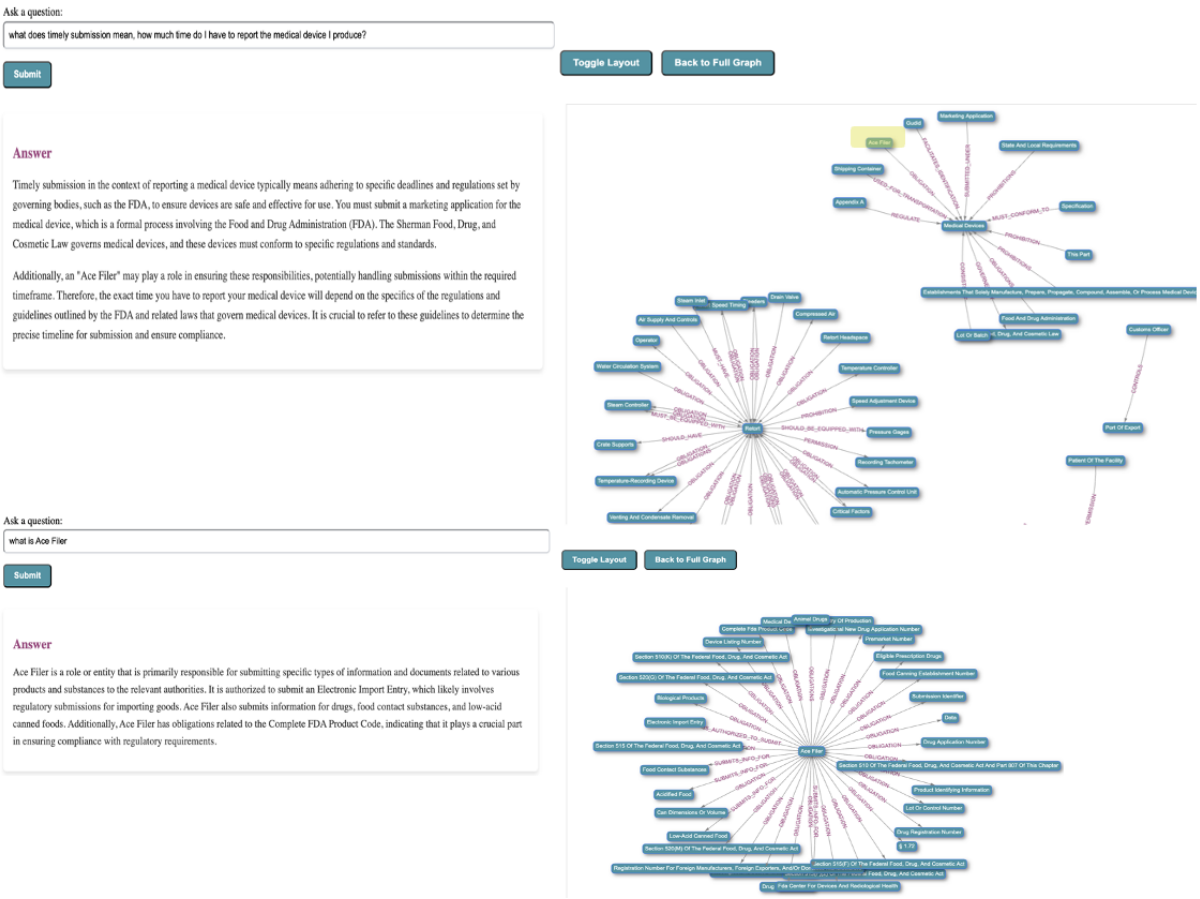

Figure 1: Navigational Facility of triplets

Figure 2: Multi-agent high-level architecture.

2.3. Triplet-first retrieval with text evidence

We compute $T_Q = \text{TopK}(\text{sim}(e_Q, e_{t_i}))$ and recover evidence $X_Q = \bigcup_{t_i \in T_Q} \Lambda(t_i)$, then pass (Q, T_Q, X_Q) to an LLM to generate the answer A. This implements RAG [11] with structured facts to reduce hallucinations [32], and has shown utility in healthcare / pharmaceutical QA [21, 20].

2.4. Design notes

We supplement the answers with an interactive subgraph of the retrieved triplets (Figure 1) to expose how the evidence pieces connect. This improves user trust and supports auditability. The completeness / consistency of T, retrieval sufficiency, and auditable provenance are central. The schema light choice accelerates ingestion while relying on canonicalization to temper emergent vocabularies [26, 27].

3. Multi-Agent Architecture

We deploy specialized agents for ingestion, extraction, normalization/cleaning, indexing, retrieval, story-building, and generation (Figure 2). This follows established multi-agent design principles for modularity and scalability [33, 34, 35, 36] and is in line with recent regulatory KG/RAG systems [16, 17, 19, 18].

Table 1
Section overlap, answer accuracy (1–5), and navigation. Triplets help most at the higher overlap threshold (θ=0.75), and produce a more navigable graph (higher average degree, shorter paths).

Metric	Without Triplets	With Triplets
Section Overlap @ 0.50	0.0812	0.0745
Section Overlap @ 0.60	0.2700	0.2143
Section Overlap @ 0.75	0.1684	**0.2888**
Answer Accuracy (avg)	4.71	**4.73**
Avg. Degree	1.2939	**1.6080**
Avg. Shortest Path	2.0167	**1.3300**

4. Evaluation

4.1. Protocol

We sample target sections $S' \subset S$, build a ground-truth story per section by concatenating related mentions, generate Q/A with an LLM, and compare our system's retrieval and answers against these references. This mirrors open-domain QA/RAG setups [11, 31] while focusing on regulatory corpora [16, 19].

4.2. Metrics

Section-level overlap. For $G_{ij} = \{s_{ij}\} \cup M(s_{ij})$ and retrieved $R_{ij,r}$, $O(R_{ij,r}, G_{ij}) = \frac{|R_{ij,r} \cap G_{ij}|}{|R_{ij,r}|}$, optionally with a similarity threshold θ for near-matches.

Factual correctness. A secondary judge (LLM or expert) marks $a_r^\star$ consistent with the ground truth story; structured facts are expected to reduce hallucination [7, 32].

Navigation. For sections s_{ij} and $s_{m\ell} \in M(s_{ij})$, let $T(s)$ be extracted triplets. We compute $\text{Nav}(S') = \frac{1}{k}\sum_{j=1}^{k} \frac{\sum_{s_{mt} \in M(s_{ij})} |T(s_{ij}) \cap T(s_{mt})|}{\sum_{s_{mt} \in M(s_{ij})} |T(s_{ij}) \cup T(s_{mt})|}$ and graph connectivity (avg. degree, shortest path).

5. Discussion and Limitations

Schema-light design. Fast ingestion and adaptability come with vocabulary fragmentation; the emergence of selective schema plus canonicalization mitigates this [26, 27].

Extraction quality. Regulatory jargon and cross references may produce missing / noisy triplets; iterative curation and weak supervision help. Temporal/conditional logic may need rules beyond SPO.

Efficiency. Large, changing corpora benefit from incremental updates and efficient vector/graph indexing. The approach complements the domain-specific RAG work [21, 20] and the regulatory deployments of KG / RAG [16, 19, 17, 18].

6. Conclusion

A schema light KG with triplet-first retrieval and textual evidence has been successfully deployed at scale. It supports numerous compliance professionals with precise and auditable QA in regulatory domains, addressing known LLM risks [7, 8, 9, 10], with user counts rapidly expanding across regulatory teams (https://www.prnewswire.com/news-releases/mastercontrol-launches-ai-powered-regulatory-chat-to-simplify-compliance-navigation-for-life-sciences-manufacturers-302533086.html). Future work includes deeper temporal/conditional reasoning and tighter human-in-the-loop curation.

Declaration on Generative AI

During the preparation of this work, the author(s) used Overleaf AI Assist paid context-aware edits to improve grammar, spelling, word choice, and sentence structure, all specifically trained for academic writing. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] U.S. Food and Drug Administration, Fda guidance documents, 2025.

[2] J. J. Cordes, S. E. Dudley, L. Washington, Regulatory Compliance Burden, Technical Report, GW Regulatory Studies Center, 2022.

[3] Y. Han, A. Ceross, J. Bergmann, More than red tape: Exploring complexity in medical device regulatory affairs, BMJ Innovations (2024).

[4] T. Zhong, et al., Evaluation of openai o1: Opportunities and challenges of agi, 2024. ArXiv preprint.

[5] A. Yang, et al., Qwen2.5 technical report, 2024. ArXiv preprint.

[6] M. Abdin, et al., Phi-4 technical report, 2024. ArXiv preprint.

[7] Z. Ji, et al., Survey of hallucination in natural language generation, 2024. ArXiv preprint.

[8] C. Ling, et al., Domain specialization as the key to make large language models disruptive: A comprehensive survey, 2024. ArXiv preprint.

[9] D. Wang, S. Zhang, Large language models in medical and healthcare fields: Applications, advances, and challenges, Artificial Intelligence Review (2024).

[10] J. B. Hakim, et al., The need for guardrails with large language models in medical safety-critical settings: An artificial intelligence application in the pharmacovigilance ecosystem, 2024. ArXiv preprint.

[11] P. Lewis, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. ArXiv preprint.

[12] A. Hogan, E. Blomqvist, M. Cochez, et al., Knowledge graphs, 2021. ArXiv preprint.

[13] A. Hogan, E. Blomqvist, M. Cochez, et al., Knowledge Graphs, volume 12 of *Synthesis Lectures on Data, Semantics, and Knowledge*, Morgan & Claypool, 2021.

[14] M. Nickel, et al., A review of relational machine learning for knowledge graphs, 2015. ArXiv preprint.

[15] X. Chen, et al., A review: Knowledge reasoning over knowledge graph, Expert Systems with Applications (2020).

[16] V. Ershov, A case study for compliance as code with graphs and language models, 2023. ArXiv preprint.

[17] S. Chattoraj, et al., Semantically Rich Approach to Automating Regulations of Medical Devices, Technical Report, University of Maryland, Baltimore County, 2024.

[18] Y. Xiang, et al., Integrating knowledge graph and large language model for safety management regulatory texts, in: Lecture Notes in Computer Science, volume 14250, 2025, pp. 976–988.

[19] L. Hillebrand, et al., Advancing risk and quality assurance: A rag chatbot for improved regulatory compliance, 2024. Available on IEEE Xplore.

[20] J. Kim, M. Min, From rag to qa-rag: Integrating generative ai for pharmaceutical regulatory compliance process, 2024. ArXiv preprint.

[21] R. Yang, et al., Retrieval-augmented generation for generative artificial intelligence in health care, npj Digital Medicine (2025).

[22] O. Etzioni, A. Fader, J. Christensen, S. Soderland, Mausam, Open information extraction: The second generation, in: Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI), 2011, pp. 3–10.

[23] A. Fader, S. Soderland, O. Etzioni, Open information extraction for the web, Communications of the ACM 57 (2014) 80–86.

[24] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr., T. M. Mitchell, Toward an architecture for never-ending language learning, in: Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI), 2010, pp. 1306–1313.
[25] S. Riedel, L. Yao, A. McCallum, Relation extraction with matrix factorization and universal schemas, in: Proceedings of NAACL-HLT, 2013, pp. 74–84.
[26] L. Galárraga, C. Teflioudi, K. Hose, F. M. Suchanek, Canonicalizing open knowledge bases, in: Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM), 2014, pp. 1679–1688.
[27] W. Shen, J. Wang, P. Luo, M. Wang, A survey on entity linking: Methods, techniques, and applications, IEEE Transactions on Knowledge and Data Engineering 27 (2014) 443–460.
[28] F. Probst, S. Eck, W. Kuhn, Scalable semantics: A case study on ontology-driven geographic information integration, International Journal of Geographical Information Science 20 (2006) 563–583.
[29] J. Lehmann, et al., Dbpedia – a large-scale, multilingual knowledge base extracted from wikipedia, Semantic Web (2015).
[30] F. M. Suchanek, G. Kasneci, G. Weikum, Yago: A core of semantic knowledge unifying wikipedia and wordnet, in: Proceedings of the 16th International Conference on World Wide Web (WWW), 2007, pp. 697–706.
[31] V. Karpukhin, et al., Dense passage retrieval for open-domain question answering, 2020. ArXiv preprint.
[32] J. Li, et al., Enhancing llm factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases, 2024. ArXiv preprint.
[33] Y. Shoham, K. Leyton-Brown, Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations, Cambridge University Press, 2008. URL: https://www.eecs.harvard.edu/cs286r/courses/fall08/files/SLB.pdf.
[34] M. Wooldridge, An Introduction to MultiAgent Systems, 2 ed., Wiley, 2009.
[35] G. Weiss, Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence, MIT Press, 2000. URL: https://ieeexplore.ieee.org/book/6267355.
[36] A. Zygmunt, et al., Agent-based environment for knowledge integration, 2013. ArXiv preprint.