Knowledge Graph Enhanced Enterprise RAG Framework for the Insurance Industry

Jiong He[a], **Zac Wong**[b], **Yangtao Wang**[c], **Xiangang Cheng**[c], **Yuhui Yao**[c], **Yao Chen**[a], **Bingsheng He**[a]

[a] *School of Computing, National University of Singapore* hejiong@nus.edu.sg, yaochen@nus.edu.sg, hebs@comp.nus.edu.sg,

[b] *Asia Institute of Digital Finance, National University of Singapore* zac_wong@nus.edu.sg

[c] *FWD Group* bob.wang@fwd.com, charles.cheng@fwd.com, yuhui.yao@fwd.com,

1. Introduction

The increasing complexity of the insurance and financial sectors has increased the demand for intelligent knowledge retrieval and reasoning. AI-driven solutions in these industries must ensure accuracy, relevance and compliance with regulatory and business requirements. This paper presents an enterprise-level Retrieval Augmented Generation (RAG) framework tailored for financial applications. Our system is designed to support key industry use cases, such as agent training, recruitment, and decision support, ensuring high contextual precision while maintaining scalability and interpretability.

2. Technical Design

Our framework integrates Knowledge Graph (KG), Hypothetical Document Embedding (HyDE), and various optimizations to enhance retrieval-augmented strategies. By combining structured domain knowledge with AI-driven adaptive retrieval mechanisms, the system significantly improves response accuracy and relevance while mitigating hallucinations. Initial evaluations on agent exam test datasets (three modules from SG CMFAS) demonstrate promising results, paving the way for advanced AI adoption in the insurance sector. As shown in Fig. 1, the system architecture consists of two main components:

- **Execution Layer:** This layer optimizes the preprocessing of incoming queries. Techniques such as HyDE and query expansion generate intermediate queries with broader semantic coverage, thus improving the retrieval of relevant document fragments. The preprocessed queries are used to retrieve relevant content from the knowledge store, which employs hybrid document storage formats.
- **Knowledge Store:** The knowledge store maintains multiple representations of original documents processed in different ways. A vector database stores documents chunked using a semantic-completeness-preserving method [1]. In addition, documents are parsed and structured into a knowledge graph built on LightRAG [2].

Table 1 presents the experimental results for three methods: Base [3], Base KG [4, 5, 2], and Ours. The Base and Base KG methods utilize conventional vector databases and knowledge graphs without domain-specific optimizations or query preprocessing. Our optimized RAG framework achieves up to a 47% improvement over the Base method and up to a 20% improvement over Base KG, demonstrating the effectiveness of our approach.

Table 1: Performance comparison of different methods on SG CMFAS exam datasets.

Modules	Base	Base KG	Ours
RES5	64.29	68.49	82.14
M9M9A	65.91	77.39	76.96
HI	54.35	79.13	80.00

3. Related Work

Retrieval Augmented Generation (RAG) has gained importance as a powerful paradigm to improve AI-generated responses by incorporating external knowledge sources [6]. Previous studies have explored RAG in domains such as open-domain question answering and enterprise knowledge management, using techniques such as dense retrieval and hybrid indexing. Knowledge Graphs (KG) have been widely adopted for structured reasoning and contextual understanding, while Hypothetical Document Embedding (HyDE) [7] has been proposed to improve retrieval efficiency through intermediate representations. Research has also addressed hallucination mitigation in generative models, particularly in regulated industries such as finance and healthcare. However, limited work has focused on adapting these techniques to the insurance sector, where challenges such as regulatory constraints, complex terminology, and evolving policies require specialized solutions. Our work extends these advances by integrating KG and HyDE within a customized RAG framework for insurance applications, optimizing retrieval precision, and enhancing compliance-aware AI responses.

Acknowledgments

This project is supported by FWD Group and the Asia Institute of Digital Finance (AIDF) of the National University of Singapore (NUS) through funding, hardware, data and manpower contributions.

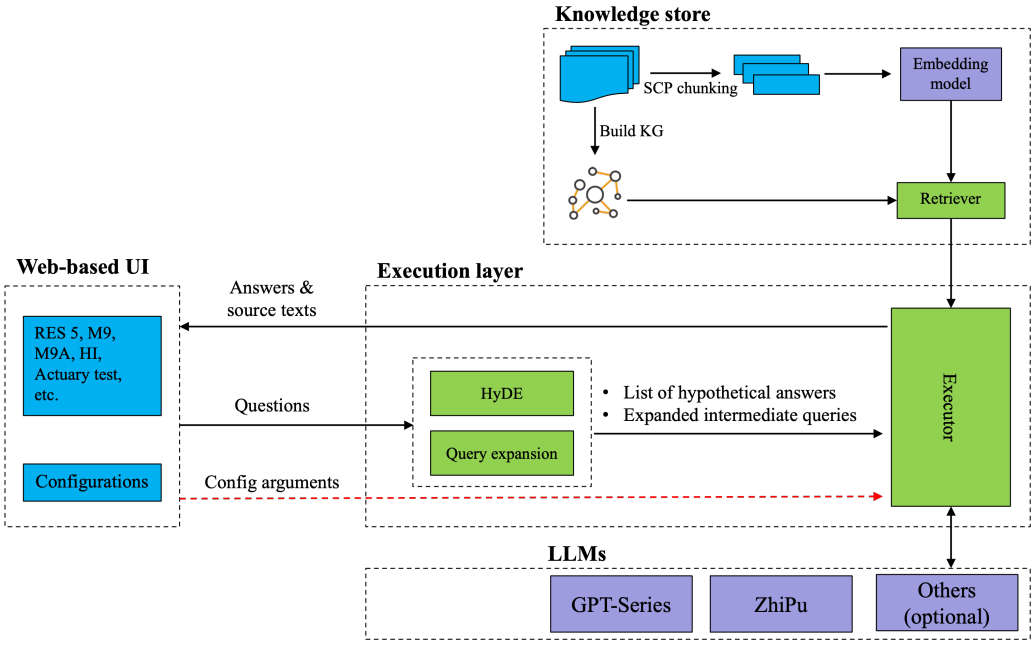

Fig. 1: Overall system architecture.

References

[1] Renyi Qu, Ruixuan Tu, and Forrest Bao. Is semantic chunking worth the computational cost?, 2024.

[2] Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. 2024.

[3] Zhi Jing, Yongye Su, and Yikun Han. When large language models meet vector databases: A survey, 2024.

[4] Bowen Zhang and Harold Soh. Extract, define, canonicalize: An llm-based framework for knowledge graph construction, 2024.

[5] Hasan Abu-Rasheed, Christian Weber, and Madjid Fathi. Knowledge graphs as context sources for llm-based explanations of learning recommendations. In *2024 IEEE Global Engineering Education Conference (EDUCON)*, pages 1–5, 2024.

[6] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 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*, KDD '24, page 6491–6501, New York, NY, USA, 2024. Association for Computing Machinery.

[7] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels, 2022.