Module 1 Capstone Milestone: Topic Approval

Alexander Ricciardi

Colorado State University Global

CSC480: Capstone Computer Science

Dr. Shaher Daoud

June 15, 2025

Module 1 Capstone Milestone: Topic Approval

A student's computer science capstone project should illustrate what the student has learned in their bachelor's program, their career goals, and specialization if any, as well as showcase their ability to solve real-world problems using software and other computer science tools. The topic that I chose to reflect my learning, career goals, and specialization (future master's in AI) is a MSHA Regulatory Conversational Agent (MRCA) web application. The MRCA is an AI-powered conversational agent that retrieves information and answers questions about Mine Safety and Health Administration (MSHA) regulations, as outlined in Title 30 of the Code of Federal Regulations (CFR). It will allow users to quickly and easily access information and ask questions about MSHA regulations using a Large Language Model (LLM) combined with a GraphRAG hybrid search.

The Business Problem

The state of Wyoming, my place of residence, has a large mining industry in the form of surface coal mines. Although I am not a miner per se, I found myself on mine sites, as mines employed an array of contractors, from heavy equipment mechanics to copier technicians like me, to support mine operations, each in their own way. To be allowed on mine sites, I had to attend a 40-hour MSHA regulations training that legally certified me as a miner, and then every year, I needed to be recertified by attending an 8-hour refresher training. These trainings are essential as the MSHA Title 30 of the CFR is a massive and complex set of regulations that is constantly changing. Furthermore, intentionally or unintentionally not following MSHA regulations on mine sites can have serious consequences for businesses and individuals. These consequences range from a couple of hundred dollars fines to ten-thousand-dollar fines, to being banned from mine sites, to jail time. Therefore, being compliant is crucial, primarily for safety

reasons, but also because MSHA agents patrol the mines and aggressively enforce the regulations.

The regulations being extensive, complex, and constantly changing create a challenge for contractors, miners, and safety managers, whose tasks need to be up-to-date and in compliance with the regulations. Even when I ensured beforehand that my task on a mine site would follow MSHA regulations, I found that once on-site, there was no easy way to check compliance if an unforeseen problem or question suddenly come to be. For example, just the printed manual "Code of Federal Regulations Title 30 Mineral Resources Parts 1 to 199 (MSHA), Revised as of July 1, 2022" from the United States Office of Federal Register (2023) has about 780 pages, and CFR Title 30 has 999 parts spread out in several printed manuals or/and sections. The sheer size and complexity of these manuals make them difficult to carry on sites and very challenging to search for information. Additionally, searching those manuals for information remains difficult even when they are saved digitally on a device. Using search engines such as Google is not a reliable method, as it is time-consuming. Similarly, using a generic LLM chatbot for fetching accurate information about MSHA regulations is not a reliable method, as they often provide just general information, but more specifically, they are prone to hallucinating, that is, they can provide plausible-sounding answers that are incorrect or entirely fabricated. Nonetheless, LLM can be incorporated into an agentic framework that utilizes a combination of vector embedding, GraphRAG (Graph Retrieval-Augmented Generation), and tool use. A Regulatory conversational agent web application can integrate these techniques, making it easily accessible (e.g., cell phone) that can not only quickly provide accurate information about MSHA regulations but also accurately explain them.

The Solution

The MSHA Regulatory Conversational Agent (MRCA) web application will be a valuable tool for mine contractors, miners, and safety managers to access accurate information about MSHA regulations quickly. The MRCA application core implementation will be a GraphRAG (Retrieval-Augmented Generation) system utilizing a hybrid search approach that combines native vector search and traversal search. A GraphRAG is a RAG technique based on a knowledge graph. A knowledge graph is a structure that stores entities and their relationships, that is, contextual data. The graph's contextual data can be queried and retrieved by an LLM, enabling it to generate accurate results based on entities and their relationships. the GraphRAG can use a hybrid search approach; it can use vector embedding for native vector search and knowledge graphs for entity and traversal search (context search). The GraphRAG hybrid search method has a Mean Reciprocal Rank (MRR) -mean accuracy- of 0.927, which is an improvement of 77.6% over the baseline RAG's score of 0.522 (Xu et al., 2024). A GraphRAG hybrid search method will be implemented in MRCA; however 92.70% mean accuracy score is not a 100% accuracy score; therefore, a notice should be posted that the agent is an informational tool designed to assist and guide users with MSHA regulatory questions and should not be used as a replacement for the official Title 30 CFR documentation. A Large Reasoning Model (LRM) will be utilized for generating the knowledge graphs, probably Gemini 2.5pro. The steps of building a knowledge graph using an LRM are as follows:

1. Gather the data
2. Chunk the data
3. Vectorize the data
4. Pass the data to an LLM or LRM to extract nodes and relationships

5. Use the output to generate the graph

(GraphAcademy, n.d.)

Table 1, below, illustrates the design phases/steps, components, and requirements/frameworks that will be used to build the MRCA web application.

Table 1

MRCA Web Application Requirements

	Component / Task	Frameworks / Requirements
Programing Language	Programming	Python
Data Collection	Data source	Title 30 of the Code of Federal Regulations (CFR) from GovInfo.gov, PDF data, for one-year bulk download. eCFR website for daily updates.
	Source data storage	The source MSHA regulations PDF data storage, Neo4j.
	Parsing & chunking	LangChain's `RecursiveCharacterTextSplitter` will be used to break down/chuck documents.
	Vectorization	OpenAI's `text-embedding-ada-002` model will be used to create numerical vector embeddings for text chunks to enable semantic search.
Knowledge Graph & Backend	Knowledge graph, entity & relationship extraction	LangChain's `LLMGraphTransformer` will be used with the Google Gemini 2.5pro model to extract nodes and relationships from the text chunks.
	Database (graph & vectors)	A cloud-hosted Neo4j DB (e.g., Neo4j AuraDB) will be used to store both the knowledge graph and the vector embeddings (vector store)
	Final step in creating the graph	The application will iterate through the structured output from the LLM and execute Cypher `MERGE` queries to create nodes and relationships in Neo4j.
	Backend API	A backend will be created using FastAPI to handle the chatbot's functionality via a RESTful API.

Conversational agent functionality & UI	Core component (GraphRAG) and prompt processing	The system is built with LangChain and uses `GraphCypherQAChain` to translate natural language questions into Cypher queries for the Neo4j database.
	Search	Search will be used a combining Neo4j's native vector search with graph traversal to find and explore information.
	Answer Generation	The OpenAI GPT-4o model will generate the final, human-readable answer based on the retrieved info. from the knowledge graph.
	User Interface (UI)	A user interface will be built using Streamlit and its chat elements.
Deployment	Security	API keys will be stored using environment variables and a `.streamlit/secrets.toml` file. Database access will be read-only to prevent malicious queries. After the main testing face a logging module should be implemented as well as a subscription plan.
	Deployment	The initial deployment will be to the Streamlit Community Cloud for accessibility (coworkers/friends) and testing. A containerized deployment (Docker and Docker Compose) will probably be used in the future.

Note: The table illustrates the design phases/steps, components, and requirements/frameworks that will be used to build the MRCA web application.

Project Timeline Scope

The timeline to design and implement the MRCA web application is only 8 weeks long, and when combined with attending school full time and having a job, I need to carefully define the scope of the project's initial version, so I can successfully deliver the project. For example, see Table 2, the MSHA regulation source can be accessed from two different and raw mine operational data can be accessed from the MSHA website itself. For this project, for the length of this course, I am only planning to use the data from the GovInfo.gov. However, if time allows it, and after getting authorized by the professor, I may integrate the raw operational mine data as examples of accidents, violations, and penalties to be searchable alongside the official MSHA regulations.

Table 2

Data Sources for the MRCA Project

Source	Content	Update Frequency	Legal Status	Use For	How to Download
GovInfo.gov	Official Title 30 CFR manual	Annually	Official, legally binding	Legal research, verification, and historical reference	Annual in PDF or XML formats
eCFR.gov	Unofficial current update compilation of the Title 30 CFR	Daily	Unofficial, not legally binding	For the most current version of the regulations	Web browser's "Print to PDF" or saving as HTML No direct file download
MSHA.gov	Raw mine operational data.	Daily, weekly, quarterly	Not applicable (it is not data regulation).	Research and analysis of mine accidents, violations, and penalties.	CSV or TXT format from data portals or by creating custom reports.

Note: The table illustrates the different sources for the MRCA project by providing a description of their content, update frequency, legal status, usage, and download methods.

Topic Summary

The goal of the MSHA Regulatory Conversational Agent (MRCA) web application is to allow users to quickly and easily access information about MSHA regulations by leveraging LLMs and GraphRAG hybrid search method for accuracy. Moreover, the main goal of the overall project is to illustrate what I learned throughout my Computer Science bachelor's program at Colorado State University Global (CSU Global), my career goals, and to be a bridge between my bachelor's and my future specialization in AI and Machine Learning (ML) through a Master’s in Science’s program at CSU Global. It will showcase my ability to solve real-world problems

using software and computer science techniques. The MRAC web application is a well-rounded project that will be a valuable portfolio project.

References

United States Office of Federal Register (2023, October 23). Code of Federal Regulations Title 30 Mineral Resources Parts 1 to 199 (MSHA), Revised as of July 1, 2022. *United States Office of Federal Register*. ISBN-13: ISBN-13: 979-8399878225

GraphAcademy (n.d.) *Building Knowledge Graphs with LLMs* [Online Course]. Neo4j. https://graphacademy.neo4j.com/courses/llm-knowledge-graph-construction/1-knowledge-graphs/1-knowledge-graph/

Xu, Z., Cruz, M. J., Guevara, M., Wang, T., Deshpande, M., Wang, X., & Li, Z. (2024, July 11). Retrieval-Augmented Generation with Knowledge Graphs for Customer Service Question Answering. *ACM Digital Library*, SIGIR '24: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, p.2905-2909. https://doi.org/10.1145/3626772.366137