CS 491

Project Specification Report

Team T2431

Smart Vector Query

Mehshid Atiq 22101335

Muhammad Rowaha 22101023

Ghulam Ahmed 22101001

Yassin Younis 22101310

Zahaab Khawaja 22101038

Supervisor: Uğur Doğrusöz

Innovation Expert: John Yuzdepski

Table Of Contents

1. Introduction

In today's world, the use of search engines is crucial in everyday tasks, legal frameworks, and professional workflows. However, the overwhelming amount of information available through these search engines can make it difficult to find the specific details which often get buried beneath an abundance of irrelevant results. Additionally, users of one of the most popular search engines, Google, are finding it more difficult to find the information they are looking for [1]. Smart Vector Query (svq.ai) is an AI tool designed to facilitate questions based solely on the user's predefined custom dataset. Through the use of a chatbot interface, users can get access to specific, relevant information, ensuring accurate and context-specific information retrieval. Designed with applications such as regulatory frameworks in mind, svq.ai provides easy access to critical information.

1.1 Description

Under the hood, svq.ai operates using state of the art large language model (LLM) technology alongside modern vector databases and embedding algorithms in order to store and retrieve information based on users' queries. Users start by uploading their own documents as the data source; these documents can be anything ranging from regulatory documents, codebases, or even textbooks. They are processed based on semantics and stored in a database. Users then ask questions through a chatbot interface, with their query being used to semantically search the database for the most relevant and precise answers.

1.2 High Level System Architecture & Components of Proposed Solution

One of the key aspects is the chunking and embedding process of the uploaded pdf, for instance. Since our aim is to enable the users to query large files, the system architecture of Svq.ai naturally feels distributed; a master node orchestrates chunking and embedding of large files. Based on configurations, for example, a worker node can be spinned up for every 10MBs of the input file. A task queue can be used for persistent asynchronous execution. Generated embeddings can then be returned to the master node. This way, worker nodes can be scaled based on the input file size. Moreover, since users will be uploading their documents to the cloud, data privacy is also a major concern. We aim to achieve this by storing hashed files instead of original files. Each user will be assigned a symmetric key for each datasource. If required, the file can be decrypted for the user and annotated accordingly based on query results. Cryptographic hashing of large files can be a slow process, and hence will be off loaded to another node in the system. For improving

availability, file/block replication can also be implemented (cloud services e.g. Amazom S3 support data replication out of the box). To scale inferences/queries to the vector databases, serverless functions can be used; queries will be short-lived and running a dedicated service constantly maintaining a connection-pool to the database and/or third-party resources (e.g. OpenAI API) can be costly. Using serverless functions also suits this need because users can query back-to-back and function execution environments do not shut down immediately–consequent requests will have 'warm' startups for a short period of time. Lastly, The system will have a single point of entry for all users; this node will be responsible for general user management and user authentication, and orchestration of processes using webhooks.

1.3 Constraints

Throughout the entire process of planning, building and implementing our project, it is inevitable to counter certain constraints. Some of these anticipated constraints have been explained below.

1.3.1 Implementation Constraints

During our implementation phase, an issue that will limit us is the technical difficulty of the project. As undergraduate students, it is unrealistic for us to make improvements on certain aspects of the project such as the semantic encapsulation of text, or the optimization of complex machine learning models. Therefore, we aim to use and combine existing technology in a unique way that appropriately solves the underlying problem effectively.

Svq.ai is designed to have a generic use-case, meaning anyone can create datasets as they like. This introduces some complexity as we will have to take a more liberal approach when it comes to programming and make sure the architecture is abstracted so that it can be used across multiple use cases.

Another constraint that will restrict our progress is our limited timeline. With such a large scope and given the complexity of our project, optimizing our models and making sure everything works well across multiple use cases will be a significant challenge. Svq.ai is going to be made to handle a variety of datasets across multiple use cases, meaning we will require significant testing and fine-tuning to achieve acceptable results throughout.

1.3.2 Economic Constraints

One of the primary economic constraints is the storage of users' data. For each dataset that a user creates, a new database will have to be created. Given the size of each dataset and the number of

datasets per user, the upkeep of the system can become very expensive. For this reason, it is crucial to operate efficiently in order to minimize the economic costs associated with data storage. Additionally, the business model must be well thought, limiting users' storage capacity based on their subscription roles. Moreover, our reliance on LLMs as the central part of the project makes it very resource demanding. With a need for large amounts of memory and a powerful GPU and CPU, locally running our own LLM model will be very costly [2]. An alternative is to leverage existing LLM APIs such as OpenAI. Additionally, another fundamental aspect to our project is semantic embedding, which also requires powerful hardware, also making it costly [3]. Once again, we will need to rely on existing models that are accessible through API endpoints. These endpoints will be chosen through the efficacy and cost of the respective models.

1.3.3 Ethical Constraints

With users uploading their own data as the source of their datasets, it is crucial that the data is protected and secured following industry standards. Data security is highly important given the datasets may include documents with sensitive or confidential information. This may include regulatory information, or personal records. It is important that the designed system protects this data in order to maintain user trust.

1.4 Professional and Ethical Issues

Although svq.ai has a generic use case, it is going to be optimized for the use of regulatory documents. With this context in mind, one of the major setbacks of LLMs, their tendency to hallucinate, must be kept in mind during development and evaluation. Within regulatory contexts compliance with regulations is of utmost importance because it protects businesses, customers and employees [4]. Especially in certain use cases, such as air transportation, law, or healthcare, regulatory compliance is a serious issue and it is crucial that svq.ai handles it accordingly.

1.5 Standards

During the development phase of our AI models, it is important to know whether optimizations are being made in the right direction. In order to clarify our direction of improvement, we have created an evaluation criteria and testbench that will allow us to standardize and quantify the evaluation of our models. The testbench consists of queries whose responses will be evaluated using the criteria. This will allow us to make direct comparisons between the different iterations of our system and see exactly where the differences and improvements lie. Moreover, having a quantifiable score for different aspects of the models allows us to better understand the

weaknesses and areas of improvement. The evaluation criteria can be found in appendix A and the testbench can be found in appendix B.

Furthermore, the system architecture will adhere to the following standards, just to name a few:

1. **OpenID Connect (OIDC):** this protocol will be used for user authentication and authorization. Identity and Access management providers like Keycloak support this standard protocol
2. **Advanced Message Queuing Protocol (AMQP):** Message-oriented middlewares such as RabbitMQ are based on AMQP application layer protocol and will be used for distributed computation of chunks
3. **Representational State Transfer (REST):** client/server communication will take place adhering to the RESTful standard. Data exchange format will be JSON (except for files, which will uploaded as a multipart request)
4. **Advanced Encryption Standard (AES):** encryption/decryption of file contents by the user will require cryptographic hashing based off of symmetric keys
5. **Best Matching 25 (BM25):** for querying reports/analytics of a datasource, full-text search engines like Lucene will use this standard to return results of reports' queries

2. Design Requirements

2.1 Functional Requirements

The regulatory document querying platform, svq.ai, aims to simplify interaction with regulatory documents through an advanced RAG-powered system. Here are the comprehensive functional requirements necessary for successful implementation and deployment.

2.1.1 User Interface and Navigation

The system requires an intuitive web interface that prioritizes user experience. The primary navigation structure must include clearly defined sections for Home, Reports, and Contact pages, with a persistent navigation bar across all pages. The interface necessitates prominent call-to-action elements, including "Try it now" and "Get started for free" buttons to maximize conversions. The chatbot interface must be seamlessly integrated into the main application, providing users with a natural conversation flow for document queries. The design must maintain consistent branding elements and styling throughout the platform to ensure a cohesive user experience.

2.1.2 Document Processing and Search Capabilities

At the core of the system lies the Retrieval Augmented Generation (RAG) technology implementation. This technology must process regulatory documents with high accuracy while reducing hallucination. The system must accept natural language queries from users and process them to provide accurate, contextual responses without hallucinations. The search functionality requires advanced filtering capabilities that allow users to refine their searches based on multiple parameters. The system must maintain an efficient indexing system for all uploaded documents, ensuring quick retrieval and processing of information.

2.1.3 User Management System

The platform requires a good user management system that supports both free-tier and premium access levels. User authentication and authorization mechanisms must be implemented with industry-standard security protocols. The system must track and maintain user sessions securely while storing user preferences and search history. Profile management capabilities should allow users to customize their experience and manage their document collections efficiently.

2.1.4 Regulatory Document Management System

The document management system must support various document formats commonly used for regulatory documentation. Version control functionality is essential to track document changes and updates over time. The system requires batch processing capabilities to handle multiple documents simultaneously while maintaining processing efficiency. Document security measures must be implemented to ensure the confidentiality and integrity of uploaded materials.

2.1.5 Search Results and Response Generation

The response generation system must provide context-aware answers derived directly from the source documents. Responses should include relevant citations and references to source materials. The system must support the export of search results and findings in multiple formats suitable for reporting and analysis. An audit trail system must track all queries and responses for compliance and quality assurance purposes.

2.1.6 Performance and Integration Requirements

System performance requirements dictate response times under three seconds for standard queries under normal load conditions. The platform must handle multiple concurrent users effectively without degradation in performance. Integration capabilities must include API access for enterprise clients,

allowing seamless incorporation into existing workflows. The system must maintain compatibility with major web browsers and provide responsive design for various device types.

2.1.7 Security and Compliance

Security requirements encompass comprehensive data protection measures, including encryption for data in transit and at rest. The system must comply with relevant data protection regulations and implement appropriate data retention and deletion policies. Audit logging functionality must track system interactions for security monitoring and compliance purposes. Clear terms of service and legal disclaimers must be integrated into the platform to ensure legal compliance.

2.1.8 Reporting and Analytics

The reporting system must generate comprehensive analytics on document usage, query patterns, and system performance. Reports must be available in multiple formats and support custom report generation based on user requirements. Analytics functionality should provide insights into user behavior and system utilization patterns to support continuous improvement.

2.2 Non-Functional Requirements

2.2.1 Usability

To guarantee that users can easily find and use functions, the interface must have clear and simple navigation. With just three clicks from any page, all of the main features should be available, reducing the amount of work needed to complete tasks. To enhance user experience and prevent misunderstanding, all pages must have a uniform style and design language. Interactivity and simplicity should be balanced in the interface so that users may finish challenging activities without feeling overburdened. It should also put aesthetics first, offering a visually pleasing layout that is enjoyable to use. All non-trivial UI elements should incorporate interactive tooltips or informational prompts to improve usability even more. This will help users navigate the system efficiently and without the need for extra assistance.

2.1.2 Reliability

The system must implement robust data security measures to ensure that all uploaded documents are protected from unauthorized access. Data confidentiality must be guaranteed through end-to-end encryption, with all data stored in an encrypted format using industry-standard algorithms (e.g., AES-256). During transmission, TLS (Transport Layer Security) protocols must be employed to safeguard data against interception or tampering. Access control mechanisms, including role-based access control (RBAC) and multi-factor authentication (MFA), must be enforced to limit access to authorized

users only. Additionally, zero-knowledge architecture can be utilized to ensure that data is not perceivable even by the system administrators or service providers. Regular penetration testing should be conducted to identify and mitigate potential vulnerabilities, ensuring data confidentiality at all times. The system must ensure a minimum uptime of 95%. Robust error-handling mechanisms must enable recovery from failures without data loss. Automated data backups must occur every 7 days, with redundancies in place to safeguard critical information. Rigorous testing, including stress testing, is required to ensure stability and reliability under varying conditions, maintaining user trust and consistent performance in high-stakes applications.

2.1.3 Performance

The system must ensure consistent performance, tailored to the app's specific functionalities. Response times for standard interface actions should not exceed 5 seconds, while complete responses must be processed within 10 seconds. Notifications should be dispatched within 3 seconds to ensure timely communication with users. File uploads must initiate within 20 seconds, and real-time information transfer via WebSockets must have a maximum lag of 2 seconds. Efficient resource utilization, optimized handling of vector embeddings and large language models, caching mechanisms, and load balancing must be implemented to support seamless scalability and responsiveness, even under heavy usage or complex queries. These measures will ensure a reliable and efficient experience for users engaging with large datasets and regulatory documents.

2.1.4 Supportability

The system must include user manuals updated with each release to guide users on app usage and features. Maintenance checks must occur regularly, with seamless, backward-compatible updates to minimize disruption. Users must be notified 1 day before scheduled updates, and downtime should be kept minimal. Comprehensive documentation for developers and administrators must cover architecture, APIs, and troubleshooting. A robust support system with FAQs, and a knowledge base must ensure timely issue resolution and user satisfaction. Monitoring and automated alerts should promptly detect and address potential issues.

2.1.5 Scalability

The system must support horizontal scalability by allowing the addition of servers and elastic storage to dynamically adjust based on user and data demands. Vertical scalability should enable scaling of system components, such as CPU and memory, to maintain performance standards during high usage. The architecture must be designed to efficiently handle growing user bases and datasets without compromising

response times or reliability. Regular stress testing should ensure the platform's ability to scale seamlessly as requirements evolve.

3. Feasibility Discussions

This section will discuss the feasibility of Svq.ai by highlighting the market potential and competitive landscape it will operate within. Additionally it will discuss the project's feasibility from an academic perspective.

3.1 Market and Competitive Analysis

Svq.ai aims to address the lack of highly relevant, accurate, reliable querying on personalized data, this data can include image, video, audio, and text. The software will target the average consumer, who may not have the best technical skills, and will have a simple and intuitive UI.

The market for such a solution will be vast, as it can speed up business processes for professionals from all different fields and can assist in daily tasks for casual users, like students. Potential use cases are outlined below:

3.1.1 Personalized Study Tool

Problem: Students often find it challenging to locate specific topics or concepts within textbooks or lecture notes. Additionally when querying for information students are often presented with results that don't align with their syllabus.

svq.ai's Solution:

- Student uploads school textbook and teacher lecture notes.
- Student queries *"How to solve for Fourier Transforms"*.
- Student receives an answer which outlines how to solve Fourier Transforms with a reference to where it was found in the textbook or lecture notes.

3.1.2 Regulatory and Legal Applications

Problem: Legal professionals and compliance officers often navigate complex regulatory documents and case law to find specific clauses or rulings. Existing tools are expensive, require expertise, and fail to cater to natural language queries. Non-lawyers in fields like finance or healthcare also struggle with understanding regulatory requirements.

svq.ai's Solution:

- A lawyer uploads regulatory documents such as GDPR or case law.
- The lawyer queries, "What are the privacy regulations for digital marketing under GDPR?"
- svq.ai provides a concise answer referencing the relevant articles and clauses within GDPR. For example: *"Article 6(1)(f) specifies conditions for lawful data processing."*

3.1.3 Academic Research

Problem: Researchers spend hours sifting through journals, textbooks, or datasets to find specific data points or summarized insights. Irrelevant results often hinder the research process.

svq.ai's Solution:

- A researcher uploads specific articles on renewable energy advancements.
- The researcher queries, *"What is the numerical result for the P-test in the data?"*
- svq.ai returns key points from these specific articles, along with citations and links to the exact sections in the uploaded documents.

Svq.ai's use cases are more vast and can be used for much more than this, indicating the existence of a massive market, valued at billions of dollars. Even the smallest penetration into the market would yield massive revenue.

And while some RAG solutions already exist for enterprises they normally are difficult to use and require technical expertise to handle, they handle mostly only text input as opposed to Svq.ai's more holistic approach towards data. Svq.ai promises an easy to use interface for the average consumer and targets a similar yet different market to existing solutions.

3.2 Academic Analysis

RAG models have been a recent attractive area of research, especially after the emergence of LLMs. Hundreds of research papers came out in the last month discussing the feasibility of RAG models and potential optimizations. Additionally some RAG models are available publicly for developer/academic use such as Hugging Face's open-source model [5]

During its development phase, Svq.ai will capitalize on the most recent industry findings, incorporating them into its core model. And since a lot of fine tuning of the hundreds of parameters Svq.ai will utilize

will be required, Svq.ai might even present new approaches to optimizing RAG models, contributing to the growing body of research in this field.

By staying at the forefront of RAG model advancements and maintaining a research-driven approach, Svq.ai aims not only to build a robust and competitive product but also to establish itself as a thought leader in the field of AI-driven retrieval systems.

4. Glossary

4.1 Appendix A - Evaluation Criteria

Criterion	**Sub-Criterion**	**Brief Description**	**Rating Scale (1-5)**
Accuracy	*Document Retrieval Accuracy*	How well the model retrieves the most relevant regulatory documents or sections.	1: Irrelevant documents retrieved most of the time. 2: Limited relevance. 3: Moderately relevant. 4: Mostly relevant documents retrieved. 5: Highly relevant and precise retrieval.
	Fact-Checking	Ensures generated responses align with retrieved documents.	1: Responses are mostly incorrect. 2: Frequent factual errors.

			3: Some factual errors, mostly correct. 4: Rare factual errors. 5: Fully accurate responses.
	Context Preservation	Maintains the original meaning and context of regulatory clauses.	1: Context is often lost or misinterpreted. 2: Frequent context errors. 3: Maintains context with some lapses. 4: Mostly accurate context. 5: Perfectly maintains context.
Comprehensiveness	*Regulation Coverage*	The model's ability to handle the entire scope of applicable regulations.	1: Major gaps in coverage. 2: Limited coverage. 3: Moderate coverage but with gaps. 4: Broad coverage with minor gaps. 5: Fully comprehensive.

	Handling Complex Queries	Handles nuanced, multi-part, or detailed regulatory questions effectively.	1: Fails to address complexity. 2: Struggles with complexity. 3: Handles some complexity but lacks precision. 4: Handles most complexities well. 5: Excels at addressing complex queries.
	Generic Query Handling	Ability to provide meaningful and accurate responses to broad or generic regulatory questions.	1: Provides vague or irrelevant responses. 2: Struggles to address generic questions. 3: Provides moderately useful responses but lacks depth. 4: Handles broad queries well with minor gaps. 5: Fully understands and addresses generic queries with insightful responses.

Relevance	*Query Understanding*	Ability to understand and interpret user queries accurately.	1: Often misinterprets queries. 2: Frequent misinterpretations. 3: Adequate understanding but lacks precision. 4: Strong understanding with rare issues. 5: Consistently accurate query interpretation.
	Pertinence of Outputs	Relevance of the retrieved documents and generated responses.	1: Responses are mostly irrelevant. 2: Frequently irrelevant responses. 3: Somewhat relevant. 4: Mostly relevant. 5: Highly relevant.
Legal Consistency	*Compliance with Regulations*	Ensures generated outputs align with the latest regulations.	1: Outputs are often non-compliant. 2: Frequent compliance issues.

			3: Adequate compliance with occasional lapses. 4: Mostly compliant. 5: Fully compliant with regulations.
	Consistency Across Queries	Produces consistent results for similar or identical regulatory queries.	1: Inconsistent across queries. 2: Frequently inconsistent. 3: Some inconsistencies but generally acceptable. 4: Mostly consistent. 5: Fully consistent.
Memory	*Context Retention*	The model's ability to retain information provided earlier in a conversation.	1: Forgets context frequently. 2: Retains context inconsistently. 3: Retains context moderately well but may miss key details. 4: Retains context with minor lapses.

			5: Excellent retention of context across multi-turn conversations.
	Memory Accuracy	How accurately the model remembers and applies past context to new queries.	1: Frequently recalls inaccurate or irrelevant information. 2: Often applies context incorrectly. 3: Moderately accurate recall. 4: Accurate recall with rare issues. 5: Consistently accurate and contextually relevant recall.

4.2 Appendix B - Testbench

Document Retrieval Accuracy:

1) What are the licensing requirements for a commercial pilot in Canada according to Transport Canada regulations?
2) What are the rules and restrictions for flying drones near airports in Canada?
3) What airworthiness standards must be met for operating small aircraft in Canada?
4) What regulations govern aircraft noise levels and emissions in Canada?

5) What are the regulatory requirements for emergency procedures on Canadian commercial flights?

Fact Checking:

1) What are the rights of passengers if their flight is delayed or canceled in Canada according to the Air Passenger Protection Regulations?
2) What are the legal requirements for maintaining aircraft maintenance records in Canada?
3) What are the maximum allowable flight duty hours for commercial pilots in Canada?
4) What are the operational requirements for flying in controlled versus uncontrolled airspace in Canada?
5) What are the rules for transporting dangerous goods by air in Canada?

Context Preservation:

1) What are the medical certification requirements for private pilots in Canada?
2) What are the minimum equipment requirements for operating under Instrument Flight Rules in Canada?
3) What regulations exist in Canada for managing wildlife hazards at airports?
4) What are the conditions an organization must meet to obtain an air operator certificate in Canada?
5) What are the minimum altitude requirements for aircraft flying over populated areas in Canada?

Regulation Coverage:

1) Outline the requirements for a Canadian-registered aircraft flying to the United States, including customs, documentation, and airspace regulations.
2) Explain the radio communication protocols required for flights operating in controlled airspace in Canada.
3) Discuss the security screening requirements for passengers boarding domestic flights in Canada.
4) Are Canadian air operators required to report their greenhouse gas emissions? If so, what are the procedures?

5) Under what circumstances can small aircraft operate at night in Canada?

Handling Complex Queries:

1) What are the regulatory requirements for operating a commercial flight in Canada, including pilot licensing, aircraft maintenance, and passenger safety protocols?
2) How do the regulations differ for an aircraft transitioning from uncontrolled to controlled airspace in Canada, and what are the communication and equipment requirements in both cases?
3) What steps must a drone operator take to legally fly in a restricted area in Canada, and how do these requirements vary based on the weight of the drone?
4) What are the requirements for Canadian airlines to comply with environmental regulations, including noise restrictions, fuel efficiency, and carbon offset programs?
5) What are the legal obligations of a flight training school in Canada regarding student safety, aircraft maintenance, and reporting incidents to Transport Canada?

Generic Query Handling:

1) What are the main categories of regulations governing aviation in Canada?
2) What topics are covered in Canadian air regulations?
3) How is the Canadian Aviation Regulations document organized?
4) Who are the key stakeholders impacted by Canadian air regulations, and how are they addressed in the document?
5) What is the primary purpose of Canadian aviation regulations?

Query Understanding:

1) Why do Canadian aviation regulations exist?
2) What do Canadian air operators need to comply with?
3) Have there been recent updates to Canadian aviation laws?
4) How do Canadian regulations ensure safety in aviation?

5) Do these regulations apply to all types of aircraft?

Pertinence of Outputs:

1) What rules ensure passenger safety during commercial flights in Canada?
2) Which regulations outline the maintenance standards for Canadian-registered aircraft?
3) What compliance requirements must drone operators in Canada meet?
4) What are the licensing requirements for pilots flying in Canada?
5) How do Canadian air regulations address environmental concerns?

Compliance with Regulations:

1) What are the current age restrictions for pilots operating commercial flights in Canada?
2) What are the most recent regulations for operating drones in Canada?
3) Have there been any new requirements for airlines in Canada to report environmental data?
4) What recent updates have been made to air traffic control procedures in Canada?
5) What are the latest certification standards for new aircraft in Canada?

Consistency Across Queries:

1) Pair 1:
    a) What are the licensing requirements for commercial pilots in Canada?
    b) What qualifications are needed to become a commercial pilot in Canada?
2) Pair 2:
    a) What are the operational rules for drones in Canadian airspace?
    b) How are drones regulated in Canada?
3) Pair 3:
    a) What are the maintenance requirements for Canadian-registered aircraft?

b) How often must Canadian aircraft undergo maintenance checks?

4) Pair 4:

a) How do Canadian regulations address aviation emissions?

b) What are the rules for controlling greenhouse gas emissions from aviation in Canada?

5) Pair 5:

a) What safety measures are required for passengers during Canadian domestic flights?

b) How do Canadian air regulations ensure passenger safety on domestic flights?

Context Retention:

1) Scenario 1:

a) What are the general requirements for operating aircraft in Canada?

b) Do these requirements apply to drones as well?

2) Scenario 2:

a) What safety measures are required for passengers on Canadian domestic flights?

b) Are these safety measures different for international flights?

c) Considering the safety measures required for domestic and international flights, what additional steps are outlined for passenger safety during emergency landings?

d) How do these safety measures reflect the overall purpose of Canadian aviation regulations?

Memory Accuracy:

1) Scenario 1:

a) What are the general maintenance requirements for Canadian-registered aircraft?

b) Based on these requirements, how often should a commercial aircraft undergo safety inspections?

2) Scenario 2:

    a) What are the regulatory requirements for operating drones in Canadian airspace?

    b) What are the key regulatory differences between operating drones and manned aircraft in Canadian airspace?

    c) How do these regulatory differences affect the required safety procedures for drones versus manned aircraft?

    d) Considering the differences in regulations, what are the specific compliance requirements for commercial drone operators in Canada?

5. References

[1] K. Debbagh, "It's become harder to find information online," *Substack*. [Online]. Available: https://kamildebbagh.substack.com/p/its-become-harder-to-find-information?sd=pf. [Accessed: Nov. 21, 2024].

[2] M. Escobar, "Memory requirements for LLM training and inference," *Medium*. [Online]. Available: https://medium.com/@manuelescobar-dev/memory-requirements-for-llm-training-and-inference-97e4ab08091b. [Accessed: Nov. 21, 2024].

[3] "Running a SOTA 7B parameter embedding model on a single GPU," *Towards Data Science*. [Online]. Available: https://towardsdatascience.com/running-a-sota-7b-parameter-embedding-model-on-a-single-gpu-bb9b071e2238. [Accessed: Nov. 21, 2024].

[4] "Regulatory compliance," *Veritas*. [Online]. Available: https://www.veritas.com/information-center/regulatory-compliance#:~:text=Regulatory%20compliance%20is%20essential%20because,caused%20by%20non%2Dcompliant%20activities. [Accessed: Nov. 21, 2024].

[5] "RAGModel documentation," *Hugging Face*. [Online]. Available: https://huggingface.co/docs/transformers/en/model_doc/rag. [Accessed: Nov. 21, 2024].