**Azhan Saleem**

**11635219**

**ADTA 5760: Generative AI with Large Language Models**

**Final Project**

**PART I: Create a Google Colab Account**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a login page

Description automatically generated**

**PART II: Seminar: Journey of Generative AI – From Boltzmann Distribution and Markov Chain to Large Language Model**

Generative Artificial Intelligence (AI) has undergone a remarkable journey spanning over a century and a half, tracing its roots back to foundational mathematical models and evolving into sophisticated neural networks and machine learning algorithms. This essay provides a comprehensive exploration of this journey, highlighting key milestones and innovations that have shaped the landscape of generative AI.

The journey began in **1868 with Ludwig Boltzmann's** groundbreaking work on statistical mechanics and the Boltzmann Distribution. Boltzmann's theory elucidated the relationship between the energy of a system and its probability distribution across different states. He observed that states with lower energy are exponentially more likely than states with higher energy, laying the groundwork for understanding probabilistic systems. Remarkably, concepts from Boltzmann's theory, such as temperature and energy states, continue to influence the design of modern AI models, particularly Large Language Models (LLMs), where temperature settings dictate the variability and confidence of generated responses.

**Boltzmann Distribution:**

A graph of energy and energy

Description automatically generated

The early 20th century witnessed the emergence of probabilistic frameworks like the Markov Chain, pioneered by **Andrey Markov in 1906**. Markov's model provided a mathematical framework for analyzing sequences, where the probability of transitioning to the next state depends solely on the current state. Despite its simplicity, the Markov Chain laid the foundation for more complex models and algorithms in AI, particularly in the domain of natural language processing.

In **1925, Ernst Ising's Ising** **Model** introduced discrete variables representing magnetic dipole moments, which could be in one of two states. Although initially applied in statistical mechanics, the Ising Model bore a resemblance to recurrent neural networks (RNNs), albeit as a non-learning model. Its contribution lies in demonstrating the concept of interaction between adjacent units, akin to the connectivity in neural networks.

**McCulloch-Pitts Model:**

Diagram of a diagram of a function

Description automatically generated

The 20th century witnessed significant advancements with the development of artificial neurons, beginning with the **McCulloch-Pitts Model in 1943**. This model laid the groundwork for perceptrons and paved the way for neural network research. Subsequent developments, such as the **Hopfield Network in 1974**, introduced recurrent neural networks with fully connected architectures, albeit with limitations in learning fixed patterns.

The 1980s marked a significant milestone with Geoffrey Hinton and Terry Sejnowski's introduction of the Boltzmann Machine in 1983. As an Energy-Based Model (EMB), the Boltzmann Machine operated on the principle of energy minimization, akin to Boltzmann's Distribution. It represented a leap forward in generative AI, paving the way for deep belief networks and data sampling techniques like Markov Chain Monte Carlo (MCMC).

The 1990s saw the emergence of **Simple Recurrent Neural Networks** (RNNs), exemplified by the Elman Network, which addressed limitations in learning longer sequences. However, these models still grappled with issues like the vanishing gradient problem, hindering their effectiveness in capturing long-term dependencies.

The 21st century ushered in transformative advancements in generative AI, notably with the development of sequence-to-sequence models in **2014 by Ilya Sutskever**. These models revolutionized natural language processing by enabling the learning of longer sequences and leveraging parallel computation for efficiency.

**Transformer architecture:**

**A diagram of a process

Description automatically generated**

In 2017, the introduction of transformers, as articulated in the seminal paper "Attention is all you need," marked a paradigm shift in generative AI. Transformers incorporated principles of parallel computing and attention mechanisms, allowing for more efficient processing of sequences. They represent the latest technology in generative AI, powering state-of-the-art language models and facilitating breakthroughs in various AI applications.

**In conclusion**, the journey of generative AI has been marked by continuous innovation and evolution, driven by insights from statistical mechanics, probabilistic frameworks, and neural network research. Recent advancements have propelled generative AI to new heights, with transformative technologies like transformers shaping the future of AI-driven innovation across diverse domains. As investment and research in AI continue to grow, we can expect further breakthroughs and applications, ushering in a new era of intelligent systems.

**PART III: Generative AI A&Q-Search System: Planning, Requirements, Data**

**Business and Technical Requirements of the System**

The decision to develop a generative AI system stems from the recognition of its potential to significantly enhance various aspects of business operations and productivity within the chosen technology startup. The following are the key business values associated with the implementation of the generative AI system:

1. **Enhanced Knowledge Management:** The generative AI system will enable the organization to efficiently manage and access its proprietary documents, thereby streamlining knowledge retrieval processes. This will lead to improved decision-making, faster problem-solving, and enhanced collaboration among team members.
2. **Increased Productivity:** By automating content searches and providing quick and accurate answers to questions, the generative AI system will boost employee productivity. Time spent on manual information retrieval tasks can be redirected towards more strategic activities, ultimately driving business growth and innovation.
3. **Competitive Advantage**: Implementing cutting-edge AI technologies reflects the organization's commitment to innovation and staying ahead of the competition. The generative AI system will equip the startup with a powerful tool for extracting valuable insights from its data, enabling it to make informed decisions and respond swiftly to market changes.
4. **Scalability and Flexibility:** As the startup grows, the generative AI system can scale alongside the business, accommodating increasing volumes of data and evolving user needs. Its flexible architecture will ensure seamless integration with existing workflows and systems, supporting the organization's long-term growth objectives.

**Business Requirements of the Generative AI System:**

The business goals and objectives that the organization aims to achieve with the generative AI system include:

1. Improved Information Retrieval: The primary objective is to develop a system that facilitates quick and accurate retrieval of information from the organization's proprietary documents. This includes text-based searches, question-answering capabilities, and contextual understanding of user queries.
2. Knowledge Sharing and Collaboration: The system should encourage knowledge sharing and collaboration among employees by providing easy access to relevant information. It should support features such as document tagging, content recommendation, and collaborative filtering to enhance information discovery and dissemination.
3. Data Security and Compliance: Ensuring the security and confidentiality of proprietary documents is paramount. The system must comply with data protection regulations and implement robust security measures to safeguard sensitive information from unauthorized access or disclosure.

**Technical Requirements of the Generative AI System:**

1. AI Platform: The generative AI system will be developed using Google Cloud Platform (GCP) Vertex AI services, leveraging its comprehensive suite of tools and capabilities for machine learning and AI development.
2. Large Language Model (LLM): The system will utilize state-of-the-art large language models, such as GPT (Generative Pre-trained Transformer) models, for natural language understanding and generation tasks.
3. Generative AI Platform: The development will make use of popular generative AI techniques, including Retrieval Augmented Generation (RAG), Sentence Transformer, and tools provided by generative AI platforms like LangChain and Hugging Face. These platforms offer pre-trained models, fine-tuning capabilities, and APIs for seamless integration into the system.
4. Cloud Storage: The system will leverage cloud storage solutions provided by GCP for storing and accessing proprietary documents and trained model artifacts securely.
5. Vector Embeddings Generation and Management: Vector embeddings will be generated using techniques such as Word2Vec or BERT embeddings to represent documents and queries in a high-dimensional vector space. These embeddings will be managed efficiently using vector databases to enable fast and accurate similarity searches.
6. Advanced Vector Search Technologies: The system will incorporate advanced vector search technologies for efficient retrieval of relevant documents based on semantic similarity and contextual relevance.

**Data and Cloud Data Storage Requirements**

**A screenshot of a computer

Description automatically generated**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* Created two buckets on Google Cloud Platform:
  + BUCKET 1: Name = adta5760-docs-folder-11
  + BUCKET 2: Name = adta5760-docs-folder-12
* Captured screenshots of both buckets:
  + Bucket 1: adta5760-docs-folder-11
  + Bucket 2: adta5760-docs-folder-12

**Data (PDF Documents) Requirement 1**

**A screenshot of a computer

Description automatically generated**

**Step 1: Download and Extract ZIP File**

1. Accessed the designated location to download the ZIP file named "AAA\_FOUNDATIONALS.zip."
2. Downloaded the ZIP file to the local machine.

**Step 2: Upload Contents to Cloud Folder**

1. Accessed the Google Cloud Platform Console using the provided credentials.
2. Navigated to the bucket named "adta5760-docs-folder-11."
3. Located the folder structure within the bucket and identified the path "documents/pdfs" for uploading the contents.
4. Extracted the contents of the downloaded ZIP file on the local machine.
5. Uploaded the extracted contents to the "pdfs" folder within the bucket on Google Cloud Platform.

**Data (PDF Documents) Requirement 2**

* The team selected "Technology Startups" as the topic for this task.
* Relevant keywords and phrases related to technology startups were identified to aid in the search process. Examples include "startup ecosystem," "entrepreneurship," "innovation," "venture capital," "startup success factors," and "startup funding."
* The team conducted searches in academic databases, online repositories, and credible websites using the identified keywords and phrases.
* The team verified that a minimum of 30 documents were collected, and each document met the quality standards required for the task.
* Accessed the Google Cloud Platform Console using the provided credentials.
* Navigated to the bucket named "adta5760-docs-folder-12."
* Located the folder structure within the bucket and identified the path "documents/pdfs" for uploading the contents.
* Uploaded the 30 docs within the bucket on Google Cloud Platform.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**PART IV: Generative AI A&Q-Search System: System Analysis**

**System Analysis Report: Generative AI Q&A-Search System**

1. Introduction: This project aims to develop a generative AI system that facilitates content searches, question-answering, and retrieval of information from the organization's proprietary documents. The system will leverage advanced generative AI techniques such as Retrieval Augmented Generation (RAG) and Sentence Transformer to provide accurate and contextual responses to user queries.

2. Problem Statement: The primary problem addressed by this project is the efficient retrieval of information from the organization's vast repository of documents. Employees often struggle to find relevant information quickly, leading to decreased productivity and efficiency. By developing a generative AI Q&A-search system, we aim to streamline the information retrieval process and empower users to access relevant data effortlessly.

3. System Requirements Analysis:

* Business Requirements:
  + Improve productivity by enabling quick and accurate information retrieval.
  + Enhance decision-making processes by providing timely access to relevant data.
  + Ensure the security and confidentiality of proprietary documents.
* Technical Requirements:
  + Integration with Google Cloud Platform (GCP): Vertex AI services for development and deployment.
  + Utilization of generative AI techniques such as RAG and Sentence Transformer for question-answering.
  + Efficient storage and retrieval of document embeddings using vector databases.
* Data Requirements:
  + Access to a diverse set of proprietary documents for training and testing the AI model.
  + Annotation and labeling of documents to facilitate supervised learning tasks.
  + Continuous updating of document embeddings to ensure relevance and accuracy.

4. Feasibility Analysis:

* Technical Feasibility:
  + The project is technically feasible given the availability of advanced AI tools and cloud infrastructure.
  + Technical risks include model performance degradation over time and challenges in optimizing search algorithms for large document repositories.
* Business Feasibility:
  + The project is expected to provide significant business value by improving information accessibility and employee productivity.
  + Financial risks include potential cost overruns due to extensive usage of cloud resources and unexpected delays in development.
* Operation Feasibility:
  + The system is expected to be used effectively by the organization, given its user-friendly interface and integration with existing workflows.
  + Risks include resistance to change among employees and the need for comprehensive training and support during deployment.

5. Project Management:

* Timeline:

**Phase 1: Requirements Gathering and Data Collection (1 month)**

Data Scientist: 1

Project Manager: 1

**Phase 2: Model Development and Testing (1 months)**

Data Scientist: 2

**Phase 3: Integration with GCP Vertex AI Services (1 month)**

Cloud Architect/Software Developer: 1

**Phase 4: Deployment and User Training (1 month)**

Support Specialist/Technical Trainer: 1

**Conclusion**: This project presents significant challenges but offers immense potential in enhancing information accessibility and decision-making within the organization. With proper planning, collaboration, and resource allocation, we are confident in delivering a robust and effective generative AI Q&A-search system that meets the needs of our stakeholders and contributes to the organization's success.

**PART V: Generative AI A&Q-Search System: System Design**

In designing the Retrieval Augmented Generator (RAG) system, we delved into both the high-level functional flow and the detailed design components. This system aims to seamlessly integrate information retrieval and language generation to provide accurate and contextually relevant responses to user queries. Below, we outline both the high-level and detailed design aspects of the RAG system.

**High-Level Design:**

1. **Question Ingestion:**
   * User queries or prompts are ingested into the system.
   * Queries are processed by both the retriever and the Large Language Model (LLM) generator.
2. **Information Retrieval (IR) System:**
   * External knowledge, encompassing relevant contextual information, is accessed by the retriever.
   * The retriever employs text transformation technology such as BM25, coupled with an encoder and vector search functionality.
   * Vector search utilizes Google's Approximate Nearest Neighbor Search for efficient retrieval of relevant context.
3. **LLM Generation:**
   * The ranked results from the retriever are passed on to the LLM generator.
   * The LLM, utilizing Gemini 1.0, generates text responses based on the retrieved context and the user's query.

**Detailed Design:**

1. **Ingestion:**
   * Documents containing external knowledge are ingested into the system.
   * Documents are split into chunks, which are tokenized for use by the LLM.
2. **Vector Embedding:**
   * Each chunk of text undergoes vector embedding, converting textual data into numerical vectors for efficient processing.
3. **Building Index:**
   * Vector embeddings are used to build an index, a critical step for performance optimization.
   * Streaming inserts ensure continuous updating of the index with new data.
4. **RAG Chain (Lang Chain**):
   * Multi-turn conversation flow is supported, allowing for interactive dialogue with users.
   * User questions are processed, generating query embeddings for semantic search.
   * Semantic search, employing Approximate Nearest Neighbors Search, retrieves relevant context from the vector store.
   * Retrieved document chunks are processed to identify top-k relevant documents.
   * The LLM generates responses based on the retrieved context and user queries.
   * Responses are provided to the user, completing the interaction loop.
5. **Direct LLM Response:**
   * In certain cases, user queries may bypass the retriever and directly trigger the LLM for response generation.
   * This ensures that the LLM incorporates the user's query into its response, maintaining context coherence.

**Conclusion:**

The RAG system's detailed design encompasses a robust framework for information retrieval and language generation. By integrating advanced technologies such as vector embedding, semantic search, and Approximate Nearest Neighbor Search, the system delivers accurate and contextually relevant responses to user queries, fostering seamless interaction and enhanced user experience.

**PART VI: Generative AI Q&A-Search System: System Set-Up**

* Included in the jupyter notebook.

**PART VII: Generative AI Q&A-Search System: System Development**

* Included in the jupyter notebook.

**PART VIII: Generative AI Q&A-Search System: System Testing**

* Included in the jupyter notebook.

**PART IX: Generative AI Q&A-Search System: System Clean-Up**

* Included in the jupyter notebook.

**PART X: Final Project: Final Presentation**