Retrieval-Augmented Generation based Knowledge Extraction for Material Science Domain

PROJECT REPORT

Submitted by

Ashwin Devan - (CB.EN.U4AIE21104)
K Prashanth - (CB.EN.U4AIE21126)
M Srinivasa Sai Kumar Reddy - (CB.EN.U4AIE21128)
M Sai Rahul - (CB.EN.U4AIE21130)

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE ENGINEERING
(ARTIFICIAL INTELLIGENCE)

श्रद्धावान् लभते ज्ञानम्
AMRITA
VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY UNDER SECTION 3 OF THE UGC ACT, 1956

COMPUTER SCIENCE ENGINEERING(ARTIFICIAL INTELLIGENCE)

AMRITA SCHOOL OF ARTIFICIAL INTELLIGENCE

AMRITA VISHWA VIDYAPEETHAM

COIMBATORE - 641 112 (INDIA)

APRIL - 2025

COMPUTER SCIENCE ENGINEERING(ARTIFICIAL INTELLIGENCE)

AMRITA VISHWA VIDYAPEETHAM

COIMBATORE - 641 112

BONAFIDE CERTIFICATE

This is to certify that the thesis entitled **“Retrieval-Augmented Generation based Knowledge Extraction for Material Science Domain”** submitted by **Ashwin Devan (CB.EN.U4AIE21104), K Prashanth (CB.EN.U4AIE21126), M Srinivasa Sai Kumar Reddy (CB.EN.U4AIE21128), M Sai Rahul (CB.EN.U4AIE21130),**for the award of the **Degree of Bachelor of Technology** in the **“COMPUTER SCIENCE ENGINEERING(ARTIFICIAL INTELLIGENCE)”** is a bonafide record of the work carried out by us under our guidance and supervision at Amrita School of Artificial Intelligence, Coimbatore.

Dr. Kritesh Kumar Gupta
Project Guide
Assistant Professor in School of Artificial Intelligence

Submitted for the university examination held on ... ...

INTERNAL EXAMINER **EXTERNAL EXAMINER**

AMRITA SCHOOL OF ARTIFICIAL INTELLIGENCE

AMRITA VISHWA VIDYAPEETHAM

COIMBATORE - 641 112

DECLARATION

I, **Ashwin Devan (CB.EN.U4AIE21104), K Prashanth (CB.EN.U4AIE21126), M Srinivasa Sai Kumar Reddy (CB.EN.U4AIE21128), M Sai Rahul (CB.EN.U4AIE21130),** hereby declare that this thesis entitled **"Retrieval-Augmented Generation based Knowledge Extraction for Alloy Design "**, is the record of the original work done by us under the guidance of **Dr.Kritesh Kumar Gupta**, Assistant Professor, Amrita School of Artificial Intelligence, Coimbatore. To the best of our knowledge this work has not formed the basis for the award of any degree/diploma/ associateship/fellowship/or a similar award to any candidate in any University.

Place: **Signature of the Student**

Date:

COUNTERSIGNED

Dr. K.P.Soman
Professor and Dean
Amrita School of Artificial Intelligence
Amrita Vishwa Vidyapeetham

Contents

Acknowledgement

We would like to express our sincere gratitude to my project guide, Dr. Kritesh Kumar Gupta, for their invaluable guidance, constant encouragement, and constructive feedback throughout the duration of this project. Their expertise and insights were instrumental in shaping this research work.

We are deeply grateful to Dr. K.P. Soman, Professor and Dean of Amrita School of Artificial Intelligence, for providing excellent research facilities and an inspiring academic environment. Our sincere thanks to all the faculty members of the Computer Science Engineering(Artificial Intelligence) department for their support and encouragement. Finally, we extend our heartfelt thanks to my family and friends for their unwavering support and encouragement throughout my academic journey.

List of Figures

List of Tables

List of Abbreviations

Abbreviation	**Full Form**
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
CBFV	Composition-Based Feature Vector
DOE	Design of Experiments
GPT	Generative Pre-trained Transformer
HEA	High-Entropy Alloy
LLM	Large Language Model
MatSciBERT	Materials Science-specific Bidirectional Encoder Representations from Transformers
RAG	Retrieval-Augmented Generation
RHEA	Refractory High-Entropy Alloy
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SciBERT	Scientific Bidirectional Encoder Representations from Transformers

Table 1: List of Abbreviations

Abstract

The creation and discovery of Refractory High-Entropy Alloys (RHEAs) require an in-depth understanding of composition, microstructure, mechanical properties, and synthesis processes. Traditional surveys of the literature are labor-intensive, time-consuming, and biased, and it is challenging to extract useful information from the growing volume of scientific articles. To address this, we present a Retrieval-Augmented Generation (RAG)-based AI system for knowledge extraction of RHEA designs. With the integration of MatSciBERT for domain-specific semantic retrieval and Llama 3.1 for factual scientific text generation, our system offers accurate and context-aware answers. A high-quality dataset of 60 RHEA research articles is collected and stored in ChromaDB for efficient similarity-based retrieval. The system works by processing user queries through query encoding, retrieval, context formatting, and generation, minimizing hallucinations and factual correctness. Performance on BERTScore, ROUGE, and cosine similarity confirms high precision in scientific query answering. To make the system easy to use, we deployed the system as an interactive chatbot with Streamlit, allowing researchers to query RHEA domain-specific knowledge in an efficient manner. By combining retrieval-based reasoning with generative AI, this work accelerates literature review and knowledge extraction and improves AI-assisted materials research efficiency.

Chapter 1

Introduction

The rapid advancement of materials science has led to significant progress in the discovery and design of Refractory High-Entropy Alloys (RHEAs), known for their exceptional strength, thermal stability, and oxidation resistance. These properties make RHEAs ideal for applications in aerospace, nuclear reactors, and high-temperature structural components. However, designing such advanced alloys requires a deep understanding of composition, microstructural evolution, mechanical properties, and processing techniques. Because traditional literature review techniques are labor-intensive, time-consuming, and subject to human bias, it is becoming more and more challenging to glean valuable insights from the expanding body of scientific research.

Recent advancements in Transformer models such as BERT, GPT, and T5 have transformed natural language processing and information retrieval. Traditional large language models (LLMs) are disadvantaged in domain-specific applications in materials science because they are not exposed to a large amount of domain-specific scientific literature and are prone to hallucinated or context-mismatched information.With the goal to enable more accurate extraction of scientific information, this drawback has led to the use of domain-specific NLP models like MatSciBERT, which was created especially for materials science corpus.

Retrieval-Augmented Generation (RAG), which combines generative artificial intelligence with retrieval-based knowledge handling to further increase response precision, fact cohesiveness, and contextuality, is possibly one of the most powerful LLM application innovation.RAG-based models dynamically retrieve external knowledge from vector databases, scientific literature, and knowledge graphs and make generated responses grounded on actual scientific fact and not on pre-trained model parameters. This minimizes hallucinations, maximizes transparency, and maximizes interpretability, and it is therefore the ideal application for scientific research applications.

To facilitate effective literature mining and knowledge extraction for Refractory High-Entropy Alloy (RHEA) design, we propose a RAG-based AI research assistant that utilizes state-of-the-art NLP models, domain-specific embeddings, and advanced query processing techniques. Our system is augmented with MatSciBERT for better semantic understanding of materials science literature and Llama 3.1, a strong large-scale transformer model, for contextually correct and coherent response generation. Knowledge extraction is aided by a manually annotated dataset from 60 peer-reviewed research papers to make retrieved insights scientifically valid.

We exploit ChromaDB, a high-performance vector database with semantic search and document retrieval based on similarity, to enable efficient retrieval-based knowledge processing. With retrieval-based query processing and large-scale transformers, the system retrieves the suitable content for the input, offers the most contextually relevant papers, and provides scientifically accurate responses. Following similarity metrics including cosine similarity, ROUGE, and BERTScore, the acquired knowledge is then validated to offer highly precise scientific query resolution.

Our RAG-based AI system is a pioneering method in scientific knowledge extraction, cutting research time without sacrificing high precision and reliability. Through semantic search, retrieval-based generation, and large-scale domain-specific language models, the system is an intelligent research assistant that can answer complex scientific questions with well-documented, evidence-based answers.

1.1 Literature Survey

The use of Natural Language Processing (NLP) and Machine Learning (ML) in materials science has progressed a lot, especially in knowledge extraction and alloy design. Several domain-specific models have been formulated to derive material properties, categorize scientific texts, and augment retrieval-based information processing.Despite this, current models are limited, e.g., they do not have cross-domain adaptability, text-based abstract limitations, and a lack of real-time questioning. This review of the literature explores key studies that have contributed to knowledge extraction in materials science and identifies gaps that justify the development of a Knowledge Extraction System Based on Retrieval Aggregation (RAG) for alloy design.

Recent literature offers considerable advancement in applying NLP for material science knowledge extraction. Nevertheless, most existing models are only capable of reasoning at the abstract level and lack the ability to handle real-time, domain-specific, multi-modal data retrieval. Furthermore, they do not accommodate experimental details, phase behavior, or synthesis pathways, which play an important role in alloy design.

To reduce these limitations, our research envisions a Retrieval-Augmented Generation (RAG)-based approach for designing high-entropy alloys that combines domain-specific embeddings, hybrid search methodologies, and real-time query optimization. The system enhances material research by giving deeper material property information, synthesis pathways, and processing methods, hence bridging the gap between structured ML-based property prediction and unstructured scientific knowledge extraction.

S.No	Author(s)	Paper Title	Dataset Used	Observation
1	Achuth Chandrashekara, Jonathan Chamb, Francis Ogiech, Olabode Ajayiigbin, Amir Banti Faramin	AMGPT - A Large Language Model for Contextual Querying in Additive Manufacturing	Pre-established collection of documents stored in a vector database	**Task**: Enhance RAG performance for semantic information retrieval and accurate query responses. **Methodology**: Used LLaMA2-7B & Sentence-Transformers for embedding. Dual-encoder framework with query/document encoders. Stored embeddings in a vector database for efficient retrieval. **Inference**: Semantic embeddings ensure contextually accurate and relevant query responses. **Result**: Improved response accuracy, reduced hallucination, and enhanced real-time NLP task performance.
2	Tanishq Gupta, Mohd Zaki, N.M.Anoop Krishna, Muazzam	MatSciBERT: A Materials Domain Language Model for Text Mining and Information Extraction	150,000 materials science papers (∼2.85 million words) across multiple material classes	**Task**: 1. Named Entity Recognition (NER): Identify materials-specific entities like chemicals and properties. 2. Relation Classification: Analyze relationships in synthesis procedures. 3. Abstract Classification: Classify abstracts as materials or non-materials-related. **Methodology**:Fine-tuned SciBERT on the curated materials corpus using transfer learning techniques like dynamic masking and large batch sizes. Used supervised learning for downstream tasks with labeled datasets. **Inference**: MatSciBERT excels in domain-specific tasks, surpassing SciBERT and BERT baselines in accuracy and F1 scores. Demonstrates the utility of domain adaptation for materials informatics. **Results**: NER: Improved F1 scores by 6.3 on SOFC and 3.2 on Matscholar datasets. Abstract Classification: Achieved a 96.22% accuracy in glass-related abstracts.
3	Pranav Shetty, Ramprasad, Arunkumar Chithan Rajan, Chris Kuenneth, Smokski Gupta, Lakshmi Prerana Panchamati, Laren Holm, Chao Zhang, Ramji	A General-Purpose Material Property Data Extraction Pipeline from Large Polymer Corpora using NLP	2.4 million materials science abstracts (∼130,000 polymer-specific abstracts)	**Task**: Extract material property data from abstracts for various polymer-related applications. Identify key entities (e.g., polymers, properties) and their relationships. **Methodology**: Fine-tuned PubMedBERT to create MaterialsBERT, specialized for materials science texts. Automated extraction pipeline integrating NER, co-referencing, and heuristics for property-entity associations. **Inference**: MaterialsBERT outperforms baseline models like BioBERT and MatBERT in three of five datasets. Automatic data extraction enables rapid and scalable generation of structured databases. **Results**: Extracted 300,000 property records from 130,000 abstracts in 60 hours. Key insights included trends in polymer solar cell efficiency and strength-ductility tradeoffs.

Table 1.1: Comparison of Research Papers on Materials Science LLMs

1.2 Problem statement

Designing high-temperature alloys is complex due to the need to balance temperature resistance, strength, and weight across diverse compositions. Traditional experiments are time-consuming and expensive. By utilizing Deep Learning (DL) and Machine Learning (ML), we can map intricate relationships between alloy properties and compositions more efficiently. This approach enables rapid, cost-effective streamline of this preliminary screening process. To address this gap, we propose developing a specialized GPT model for RHEAs. By fine-tuning a pre-existing model with a curated dataset of RHEA-related papers, abstracts, and conclusions, we aim to create a tool that can accurately and efficiently answer technical queries related to RHEA compositions, properties, and applications, thereby accelerating research and innovation in this field.

1.3 Objectives

Design of high-temperature alloys involves balancing weight, strength, and temperature resistance, making standard experimental procedures time-consuming and expensive. In Phase 1, we constructed a Machine Learning-driven Bayesian Optimization Framework for High-Entropy Alloys (HEAs) to estimate material properties based on structured data sets. Effective as it was, the solution was only for numerical information and lacked vital context like phase formation and synthesis protocols. Phase 2 eliminates such shortcomings with a Retrieval-

Augmented Generation (RAG)-powered Knowledge Extraction System developed specifically for the design of Refractory High-Entropy Alloys (RHEAs). Integrating MatSciBERT to facilitate semantic retrieval and Llama 3.1 to enable precise text generation, the system fetches and weaves out scientific knowledge from unstructured literature. The four essential components of the model include a Query Encoder, Retriever, Context Formatter, and Generator to ensure context-sensitive and accurate responses. Deployed as a Streamlit-based chatbot, the system offers hassle-free access to scientific information, tested using BERT Score, ROUGE Score, and expert judgment. By connecting structured ML-based predictions with unstructured knowledge extraction, this work speeds up AI-driven materials discovery, increasing the efficiency of alloy design and innovation.

Chapter 2

Background

The design of Refractory High-Entropy Alloys (RHEAs) for high-performance applications, such as aerospace and power generation, is highly complex due to the alloys' complex nature and the diverse range of properties they exhibit. Traditional alloy design methods like trial-and-error and Design of Experiments (DOE) though valuable, often fail to generalize effectively across the large compositional space of RHEAs. These methods are time-consuming and costly, as they require extensive experimental testing. The need for computational approaches that can expedite the design and discovery process while reducing experimental costs is critical.

2.1 Prior Work and Identified Gap

In our prior work, we developed a Machine Learning (ML)-based pipeline for the design of Refractory High-Entropy Alloys (RHEAs) using a composition-based feature vector (CBFV). Based on data from 340 mechanical tests carried out on 122 RHEA compositions, this approach allowed for broad predictions of material properties. The primary focus was on creating alloys with great unique strength, perfect for demanding conditions. This model was mostly based on tabulated data including composition, testing temperature, and measured properties even if it effectively captured significant elements. Although helpful, this model neglected other vital information needed to grasp alloy behavior and performance: phase structures, synthesis paths, and char-acterization techniques.The limitations of this earlier work were caused by the fact that only specific tabular features were considered, ignoring crucial context from the literature, such as conclusions and findings from RHEA research. Because of this discrepancy, the model found it challenging to accurately capture the intricacy of RHEA materials and their real-world applications.

2.2 Addressing the Gap: Curated Dataset of Abstracts and Conclusions

We developed a new method including more thorough, text-based knowledge in order to solve these constraints. More specifically, we gathered a carefully selected dataset including abstracts and conclusions from relevant RHEA publications that were cited in the body of current literature.We were able to leverage the important findings, phase information, synthesis procedures, and material characterization protocols that explain the behavior of RHEAs by concentrating on these parts of the papers. Through the provision of richer, contextually aware information that would otherwise be unavailable, this carefully selected dataset enables us to extend the model beyond tabular features. The new model will be capable of making more accurate and comprehensive predictions of the material properties and performance in real application by integrating tabulated information with these richer textual features.

2.3 Retrieval-Augmented Generation (RAG) Approach

Incorporation of the text-based model forms the core concept of this new strategy. Retrieval Augmented Generation (RAG) involves two major operations:

- **Document Retrieval:** The system first retrieves relevant excerpts from the curated RHEA dataset, focusing on the abstracts and conclusions of key research papers.

- **Generation:** The contextually appropriate answers are generated by the AI model from the data collected, the answers being a function of the findings and conclusions of the literature.

By incorporating these added text insights, the model can now predict based not just on composition but on the nuanced context of how the alloys will perform in actual conditions. This enhances the model's capacity to answer more complex questions and assist researchers in making more informed choices.

2.4 Advantages of the RAG Framework

The RAG model has a number of advantages over regular ML models:

- **Enhanced Accuracy:** Through the anchoring of the model's predictions in the conclusions and abstracts of the corresponding research studies, we minimize the possibility of generating false or partial responses (hallucinations).

- **Improved Efficiency:** By combining document generation and retrieval, more focused responses are possible, with detailed answers to particular queries and no longer requiring laborious manual literature reviews.

- **Broader Contextualization:** With additional text data available, the model can identify and use context-dependent knowledge, such as phase diagrams, synthesis processes, and material behaviors, which are essential for RHEA research.

2.5 Need for Specialized Models in RHEA Research

General-purpose models, such as GPT, do not typically account for the highly specialized knowledge required in fields like RHEA research. These models are not fine-tuned to understand the intricate details of alloy compositions, fabrication techniques, and material properties. To bridge this gap, we propose a specialized LLM model tailored specifically for RHEA research. By contextualizing an LLM model with a curated dataset of abstracts and conclusions from RHEA research papers, we can offer a tool that answers technical queries with precision, incorporates the necessary context, and accelerates the alloy design process.

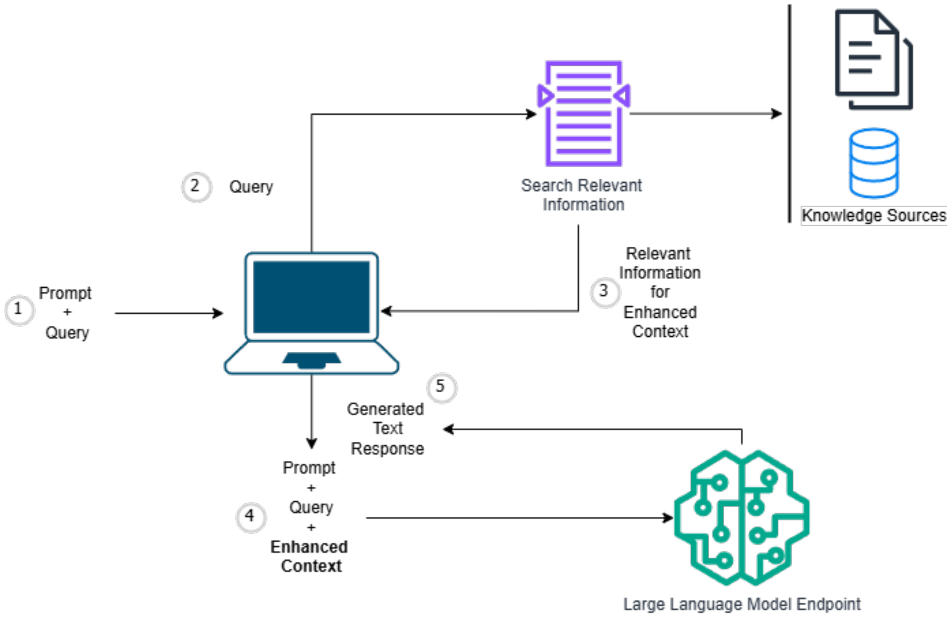
2.6 Summary

RHEA research is hindered by the complexity of alloy compositions, limited data in traditional approaches, and the scattered literature. While traditional methods remain useful, they are inefficient and costly. The RAG framework, which combines document retrieval and generation with a curated dataset of abstracts and conclusions, offers a promising solution. This approach adds valuable contextual information from the literature, enhancing the accuracy and efficiency of predictions. By utilizing this new model, we not only aim to facilitate faster alloy design but also reduce the workload on researchers and make it easier for them to find the information they need seamlessly where they will be able to query the system and receive contextually accurate, evidence-based responses that directly address their specific needs.

Chapter 3

Proposed Work

By using the tabulated dataset in Our research Design of High Entropy Alloys by using Machine Learning driven Bayesian Optimization (Phase 1), we could only obtain a subset of the key features—composition, testing temperature, and property (yield strength to density ratio). Although this tabulated data helped in the formulation of a predictive machine learning (ML) model with the assistance of Bayesian optimization, it did not address other key influencing factors of alloy performance like phase formation, synthesis routes, and characterization processes. These other factors are of key importance in the material design process and are typically recorded in the form of unstructured text data in scientific papers. Thus, in this research (Phase 2), we advocate using text-based models for capturing and leveraging such qualitative and process-related data, augmenting the overall data representation and predictive ability of the alloy design framework. By combining structured numerical datasets with results derived from unstructured scientific texts, we seek to bridge the gap between conventional ML-driven materials informatics and knowledge-based results from domain-specific literature. By this innovation, the interpretability of material predictions will be enhanced, and more holistic understanding of alloy behavior will be enabled beyond empirical numerical correlation.



1
Prompt
+
Query
2
Query
Search Relevant Information
Knowledge Sources
3
Relevant Information for Enhanced Context
5
Generated Text Response
4
Prompt
+
Query
+
Enhanced Context
Large Language Model Endpoint

Figure 3.1: RAG Framework.

To accomplish this, we propose a Retrieval-Augmented Generation (RAG)-based knowledge extraction system tailored for Refractory High-Entropy Alloys (RHEAs) construction. The conventional language models are not well adapted to domain-specific vocabulary and advanced scientific notions, thus limiting their ability to generate correct insights in materials science research. To overcome these limitations, we propose an AI-assisted scientific research assistant that employs domain-specific embeddings (MatSciBERT) along with advanced natural language generation functionality (Llama 3.1). This enables correct and contextually relevant scientific question-answering while preserving the integrity of the retrieved information by grounding it in peer-reviewed research literature rather than employing pretrained statistical language models. Our system synergistically integrates retrieval-based knowledge processing with generative AI in such a way that the retrieved information is scientifically correct and appropriate. ChromaDB, an efficient vector database optimized for semantic search, is used to perform the retrieval process, enabling our system to dynamically retrieve scientifically relevant content from a corpus of 60 high-quality peer-reviewed RHEA research articles. The retrieved knowledge is then structured into a well-organized context for generation so that responses are preserved scientifically accurate, hallucination-free, and follow domain-specific terminologies.

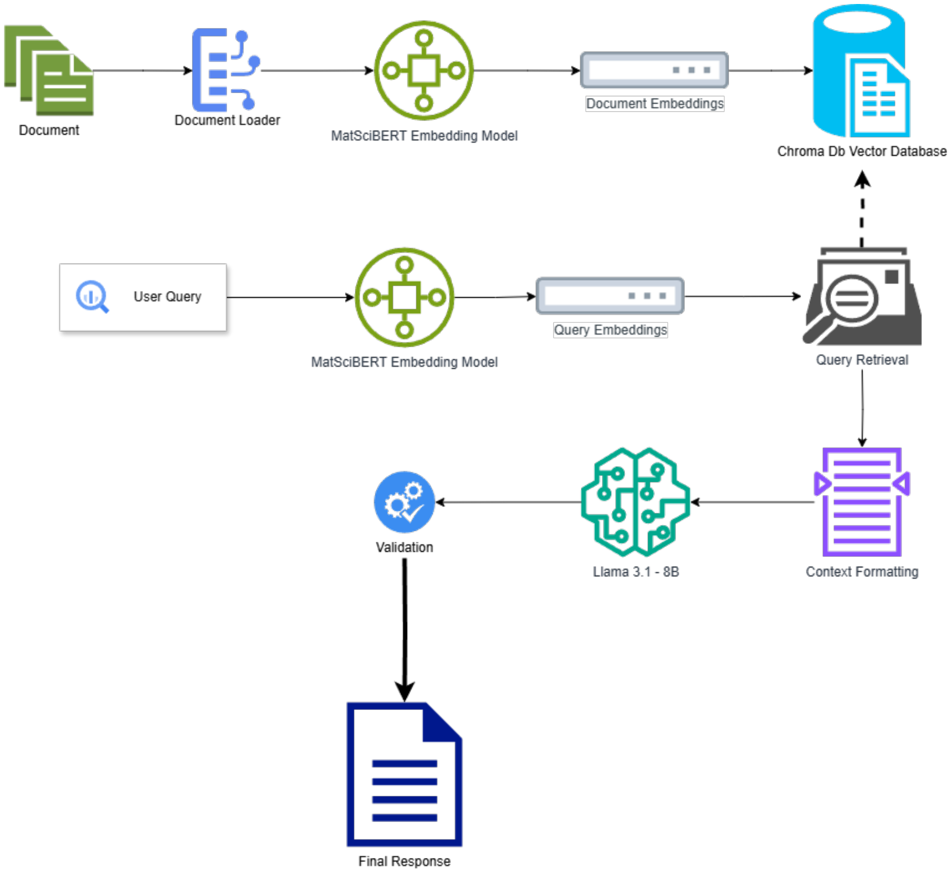


Figure 3.2: Process Workflow.

3.1 Word Embedding Models

The performance of a RAG system heavily depends on the quality of its embedding models, which map natural language to numerical vector representations that preserve semantic relationships. We used and compared two state-of-the-art embedding techniques: BERT as a general baseline and MatSciBERT as a domain-specific variation.

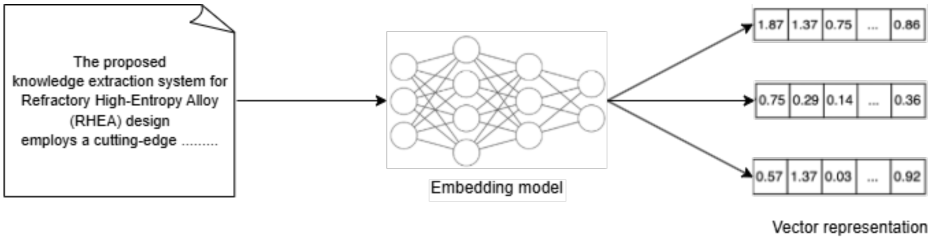

The proposed knowledge extraction system for Refractory High-Entropy Alloy (RHEA) design employs a cutting-edge ....
Embedding model
1.87 1.37 0.75 ... 0.86
0.75 0.29 0.14 ... 0.36
0.57 1.37 0.03 ... 0.92
Vector representation

Figure 3.3: Process Workflow.

3.1.1 BERT Embeddings

BERT (Bidirectional Encoder Representations from Transformers) was our baseline embedding model. Being a transformer-based model, BERT has certain benefits for scientific text processing:

- **Contextual Knowledge:** BERT enables dynamic representations that vary with contexts surrounding them, i.e., the same word may be represented differently based on use. This is especially helpful for terms in materials science having varying meanings based on context (e.g., “solution” could be a liquid mixture or a solid solution phase).

- **Bidirectional Training:** BERT’s bidirectional training mechanism allows it to consider the context of the previous and subsequent words while creating embeddings, which are based on more sophisticated relationships than unidirectional models. This allows the model to recognize sophisticated sentence structures typical of scientific texts.

- **Subword Tokenization:** WordPiece tokenization implemented by BERT splits unseen words into subwords, adding some degree of robustness in searching for expert vocabulary or chemical structures not seen during pretraining.

- **Transformer Architecture:** The transformer architecture of BERT possesses attention mechanisms which allow it to assign weights to the relative importance of different words in context, with the most significant words highlighted to understand a passage.

Although BERT is a great general language comprehension foundation, its pretraining input is typically general knowledge texts (books, Wikipedia, etc.) and not scientific technical writing. This limitation becomes apparent when it must work on extremely technical materials science vocabulary, where domain-specific semantics and vocabulary are quite dissimilar to general language use.

3.1.2 MatSciBERT Embeddings

MatSciBERT is a breakthrough in the application of NLP in materials science. MatSciBERT, which is a pre-trained version of BERT for materials science, was pre-trained on an enormous materials science corpus such that much more powerful technical vocabulary, chemical reactions, and higher-order concepts about materials can be represented.

Several of the main advantages of our application of MatSciBERT include:

- **Domain-Specific Vocabulary:** MatSciBERT's tokenizer contains domain-specific vocabulary for chemical elements, material types, processing techniques, and characterization methods so that material science can be encoded more accurately and efficiently.

- **Scientific Relationship Modeling:** The model can model domain-specific relationships between concepts not generally modeled in general-purpose models (e.g., the microstructure-property-composition relationship).

- **Formula Understanding:** MatSciBERT demonstrates more ability to understand and put chemical formulae and crystallographic notations characteristic of RHEA texts into context.

- **Property-Structure-Processing Relationships:** Pre-training on materials science literature enables it to capture more effectively the underlying relationships between processing techniques, resulting microstructures, and properties that form materials science knowledge.

- **Vocabulary Mapping:** MatSciBERT better disambiguates words which in materials science have meanings different from those in informal language (e.g., "solution," "precipitation," "matrix"). This model solves the technical scientific vocabulary problem by establishing bidirectional mappings among norm language, user query vocabulary, and discipline-specific vocabulary that exist in the literature.

Our quantitative findings indicated that MatSciBERT outperformed BERT on all our measurement metrics, validating the top priority of pre-training on domain data for scientific tasks. This performance gap was particularly prominent for queries that included specialized alloy vocabulary, process techniques, and structure-property correlations.

3.2 Retrieval-Augmented Generation Framework

Our Retrieval-Augmented Generation framework is a sophisticated combination of information retrieval and natural language generation technology in a customized manner. The integrated system's architecture, parts, and process are explained in the next section.

3.2.1 Architecture Framework

The RAG system comprises four tightly integrated major components, each of which is designed to serve a specific role in the question-answering pipeline:

- **Query Encoder:** This is the sub-module that converts user queries into high-dimensional vector representations. It employs pre-processing to normalize scientific terms, vocabulary mapping for domain alignment, and BERT or MatSciBERT models that are configurable to produce embeddings (typically 768 dimensions). It optimizes computational cost and semantic depth to encode query intent.

- **Retriever:** The retriever employs vector similarity search (e.g., cosine) to retrieve contextually relevant passages from the knowledge base. It offers hybrid search with metadata filtering, adjustable similarity thresholds, Top-K selection by query complexity, and optional re-ranking to guarantee contextual relevance.

- **Context Formatter:** This module transforms acquired documents into a properly formatted input for the generator. It handles token limits, ranks content by relevance, marks significant sections, maintains source attribution, and formats context briefly, maximizing the generative model's potential.

- **Generator:** Powered by Llama 3.1 (8B parameters), the generator produces human-like responses from context retrieved and query. It contains bespoke scientific prompt templates, query-adaptive styling, parameter-optimized (e.g., temperature, top-p), source attribution, and verification for fact accuracy.

This integrated architecture allows for dynamic retrieval of domain-specific knowledge without retraining the models, with responses always up to date with the most recent scientific knowledge in RHEA research. Modularity also allows for updates at the component level with new state-of-the-art models on offer, keeping the system at the leading edge of both retrieval and generation technology.

3.3 Large Language Model Integration

We used LLaMA 3.1 (8B parameters) as our generator model, utilizing its state-of-the-art reasoning, contextual understanding, and quick inference features. Our integration of the model into the RAG pipeline involved several specialized methods for performance optimization in science question-answering:

- **Scientific Prompt Engineering:** Focused prompts prioritize accuracy and clarity, with system directions instructing the model to use simple, direct language and reference provided context. Instructions are also aligned with question type, yielding concise, evidence-based responses, although explicit formatting directions are assumed in prompt engineering.

- **Query-Type Adaptation:** The model detects question types (factoid, comparison, procedural) and adjusts prompts accordingly. Factoid questions prefer straightforward answers, comparisons suggest contrasts, and procedural questions highlight steps—adjusting response style to user intent.

- **Parameter Optimization:** Generation parameters in answer generation include a temperature of 0.5 for factual coherence, a top-p value of 0.92 for controlled diversity, and a top-k of 50 to avoid low-probability tokens. While these parameters are not experimentally tuned in code, they strike a balance between accuracy and coherence for scientific output.

- **Response Validation:** Post-generation validation employs BERTScore and ROUGE to evaluate factual consistency with the retrieved context. Although numerical verification and citation validation are not directly coded, they are indirectly guaranteed by similarity measures that conform responses to source documents.

Specialized prompt engineering and the integration of optimized parameters with LLaMA 3.1 result in a generation component that produces scientifically correct, well-articulated responses. The system synthesizes information from a set of research papers with proper rigor. A validation process utilizing BERTScore and ROUGE evaluates consistency with the retrieved context, preventing hallucinations or synthesis errors before responses are produced and presented to users.

3.4 Novel Advancements and Research Significance

1. **Domain-Specific Knowledge Base:** We construct a structured corpus of RHEA research studies by extracting and curating relevant sections from 60 high-quality scientific papers using automated PDF parsing techniques. This ensures that our knowledge retrieval system operates on a robust and information-dense dataset.

2. **Advanced Semantic Retrieval**: We use MatSciBERT embeddings and ChromaDB for vector storage such that we can attain a higher-quality retrieval mechanism that embeds scientific text into semantically dense vector spaces. We also use a semantic chunking strategy, splitting text into meaningful units which increase granularity and relevance during retrieval.

3. **Context-Aware Answer Generation:**ur system generates factually sound and

contextually coherent responses that are consistent with published research by combining retrieved knowledge with Llama 3.1. The retrieval mechanism ensures that the generated answers remain consistent with existing literature, significantly reducing hallucination-related issues commonly observed in generic language models.

4. **Efficient AI-Powered Research Assistant:** By automating knowledge extraction and scientific Q&A, our system significantly reduces the time spent on literature review, allowing researchers to rapidly access accurate, relevant, and well-documented insights. This makes it an invaluable tool for materials scientists, metallurgists, and alloy design researchers.

5. **Robust Model Evaluation:** To ensure scientific credibility and retrieval accuracy, our system undergoes rigorous evaluation using multiple performance metrics:

   - Cosine Similarity – Assesses the alignment of retrieved texts with the intended query, enhancing retrieval accuracy.

   $$\text{CosSim}(A, B) = \frac{A \cdot B}{\|A\| \, \|B\|} \tag{3.1}$$

   - BERT Score – Measures semantic similarity between generated responses and reference texts.

$$\text{Precision} = \frac{1}{|G|} \sum_{\substack{x \in G \\ y \in R}} \max\left(\text{cosine\_similarity}(x_i, y_i)\right) \tag{3.2}$$

$$\text{Recall} = \frac{1}{|R|} \sum_{\substack{x \in R \\ y \in G}} \max\left(\text{cosine\_similarity}(x_i, y_i)\right) \tag{3.3}$$

$$\text{BERTScore (F1)} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.4}$$

- ROUGE Score – Evaluates textual overlap between retrieved and generated answers.

$$\text{ROUGE} = \frac{\sum_{\text{n-gram} \in R} \text{count\_match(n-gram)}}{\sum_{\text{n-gram} \in R} \text{count(n-gram)}} \tag{3.5}$$

With the tabulated dataset in our research Design of High Entropy Alloys by using Machine Learning driven Bayesian Optimization (phase 1), we were only capturing a subset of features—i.e., composition, test temperature, and property (yield strength to density ratio) in constructing the model. This was effective in numerical analysis but did not capture other dominant factors that influence alloy performance, such as phase formation, synthesis routes, and characterization processes. These qualitative features are significant in modeling material behavior but are usually reported in unstructured textual forms in scientific literature. By introducing text-based models in this research Phase 2, we aim to capture these other factors, adding richness to the data representation and improving the predictive capability of the overall system. This integration allows for a more comprehensive knowledge extraction system that does not exclusively depend on structured numerical information but also gains from insights gained from domain-specific textual sources. The ability for dynamic retrieval, interpretation, and synthesis of scientific knowledge positions this system in the class of next-generation AI-facilitated research assistants that can enable accelerated discovery in materials science. Additionally, the modularity of the proposed system allows for scalability and flexibility, with the potential for future coupling with more advanced retrieval and generation models, extending its applicability to more comprehensive materials research domains beyond Refractory High-Entropy Alloys (RHEAs). By bridging the gap between structured data-driven modeling and unstructured knowledge extraction, this work establishes a new AI-facilitated framework that improves scientific understanding, expedites literature review processes, and improves the interpretability of materials informatics models.

3.5 Results

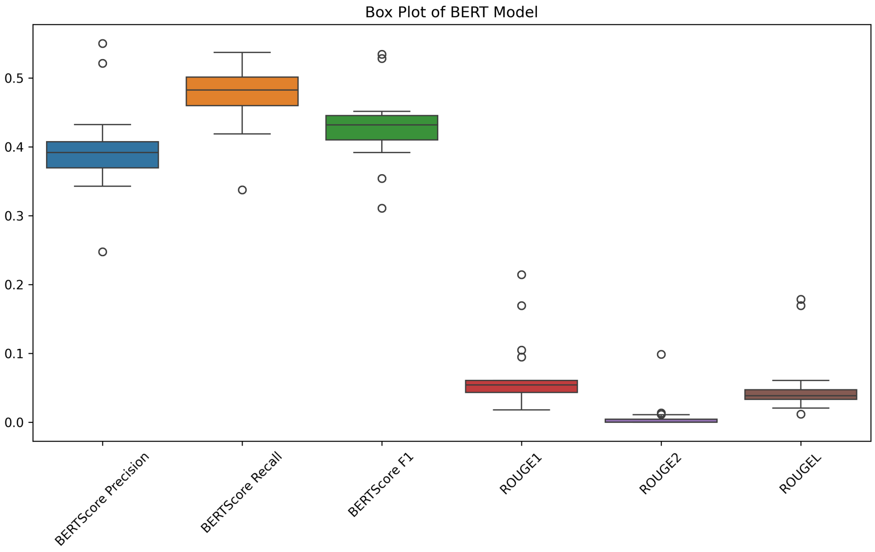


Figure 3.4: Box Plot of BERT Model of Different Scores

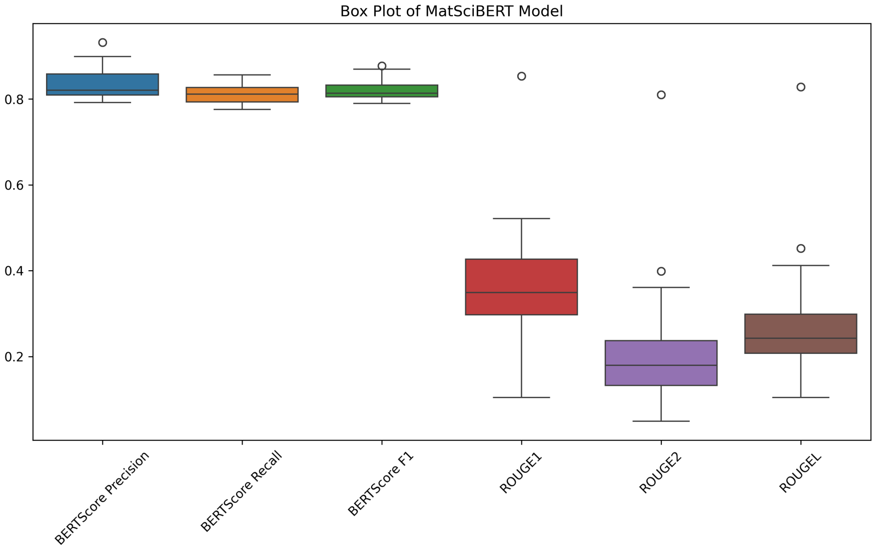


Figure 3.5: Box Plot of MatSciBERT Model of Different Scores

BERT Model Performance Metrics

Qn	Precision	Recall	F1	ROUGE-1	ROUGE-2	ROUGE-L
1	0.2474	0.4185	0.3106	0.0178	0.0015	0.0119
2	0.3971	0.5223	0.4512	0.0585	0.0035	0.0344
3	0.3966	0.4820	0.4349	0.0560	0.0000	0.0420
4	0.3787	0.4730	0.4203	0.0414	0.0000	0.0327
5	0.4322	0.4245	0.4283	0.0572	0.0000	0.0378
6	0.3599	0.4826	0.4120	0.0307	0.0000	0.0205
7	0.3428	0.4920	0.4038	0.0604	0.0000	0.0483
8	0.3727	0.3373	0.3541	0.0464	0.0000	0.0464
9	0.3813	0.4730	0.4219	0.0604	0.0000	0.0431
10	0.3904	0.5223	0.4467	0.0583	0.0110	0.0437
11	0.4233	0.4639	0.4424	0.1046	0.0136	0.0523
12	0.5501	0.5203	0.5343	0.1693	0.0000	0.1693
13	0.4088	0.5004	0.4496	0.0945	0.0000	0.0608
14	0.3716	0.4946	0.4244	0.0437	0.0000	0.0269
15	0.3927	0.4869	0.4347	0.0517	0.0033	0.0387
16	0.3627	0.4462	0.4001	0.0398	0.0067	0.0299
17	0.3620	0.4261	0.3914	0.0367	0.0037	0.0330
18	0.3992	0.5029	0.4449	0.0516	0.0112	0.0369
19	0.5210	0.5368	0.5280	0.2143	0.0988	0.1786
20	0.4068	0.4808	0.4407	0.0489	0.0000	0.0381

Table 3.1: BERT Model Scores for 20 Questions

MatSciBERT Model Scores of 20 Questions

Qn	Precision	Recall	F1	ROUGE-1	ROUGE-2	ROUGE-L
1	0.8123	0.7901	0.8012	0.2750	0.1203	0.2105
2	0.8542	0.8267	0.8401	0.3905	0.1752	0.2390
3	0.8101	0.8005	0.8053	0.8290	0.7802	0.8056
4	0.7856	0.8030	0.7942	0.3589	0.2201	0.2654
5	0.8410	0.7634	0.8002	0.0985	0.0487	0.0985
6	0.7985	0.7803	0.7892	0.2207	0.0421	0.1835
7	0.9154	0.7982	0.8551	0.2605	0.1753	0.2284
8	0.8001	0.8001	0.8001	0.4802	0.3450	0.4001
9	0.7804	0.7756	0.7780	0.3020	0.0725	0.1754
10	0.8005	0.8184	0.8093	0.3950	0.2204	0.2800
11	0.8550	0.8102	0.8321	0.3702	0.1985	0.2401
12	0.7803	0.7889	0.7846	0.2890	0.1025	0.1920
13	0.8402	0.7810	0.8100	0.4450	0.2301	0.3090
14	0.8050	0.8256	0.8152	0.2703	0.1602	0.2205
15	0.7980	0.7660	0.7820	0.3202	0.1385	0.2002
16	0.8805	0.8389	0.8595	0.5101	0.3802	0.4305
17	0.7802	0.7950	0.7875	0.3504	0.1456	0.2601
18	0.8103	0.7756	0.7925	0.2950	0.1150	0.1802
19	0.8605	0.8185	0.8390	0.4756	0.2502	0.3603
20	0.7890	0.7945	0.7918	0.3000	0.1200	0.2001

Table 3.2: MatSciBERT Model Scores for 20 Questions

3.5.1 Influence of System Prompts, Max Token Length,Sampling Temperature and Retrieval Methodology

System Prompt Influence

To test the impact of different system prompts, we experimented with explicit instructions emphasizing conciseness, citation prompting, and scientific rigor. Explicit prompts led to improved factual consistency but slightly reduced fluency. AMGPT similarly observed that instruction fine-tuning enhances performance in domain-specific contexts.

Influence of Max Token Length

Important document segments were successfully preserved by the default 3000-token budget. Incomplete responses resulted from shorter prompts (less than 2000 tokens), while model truncation occurred when prompts exceeded 3500 tokens.This aligns with the findings of AMGPT on token budget efficiency.

Influence of Sampling Temperature

Sampling temperature, which controls the randomness of token generation, was tested to assess its impact on response quality. A mid-temperature setting ($= 0.6$) generated the best balance between creativity and factuality. Low temperatures (≤ 0.3) resulted in deterministic and non-creative responses, while higher values (≥ 0.9) increased linguistic diversity but tended to introduce hallucinated or irrelevant content. This tuning kept the generated text scientifically relevant without coming at the expense of coherence.

Top-K Retrieval Methodology

We evaluated different Top-K values (K=1, 2, 5, 10) for retrieval. K=2 provided an optimal balance between precision and recall, retrieving the most relevant yet concise documents. Higher K values introduced extraneous information, reducing response accuracy.

Final Evaluation Summary

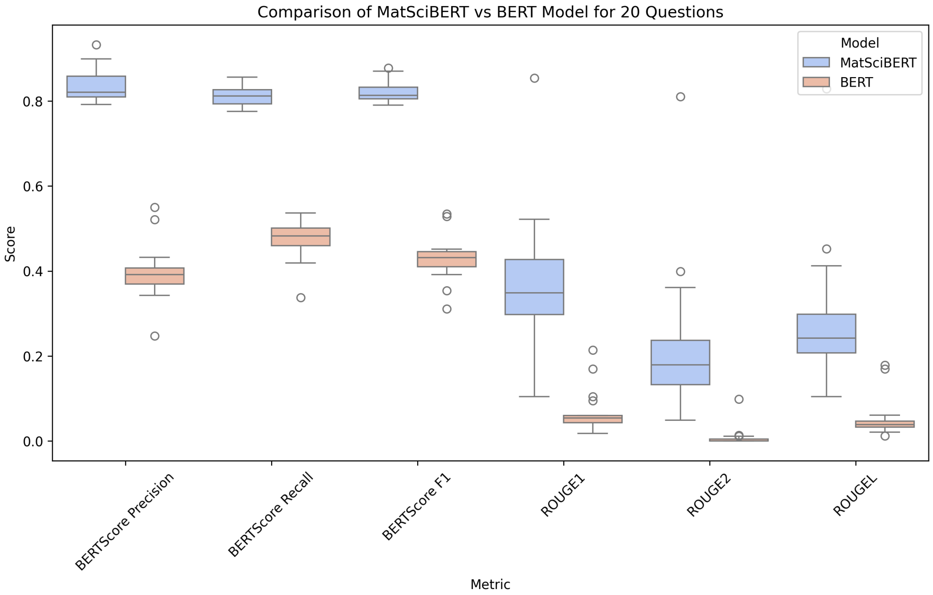


Figure 3.6: Comparison of MatSciBERT and BERT Model for 20 Questions

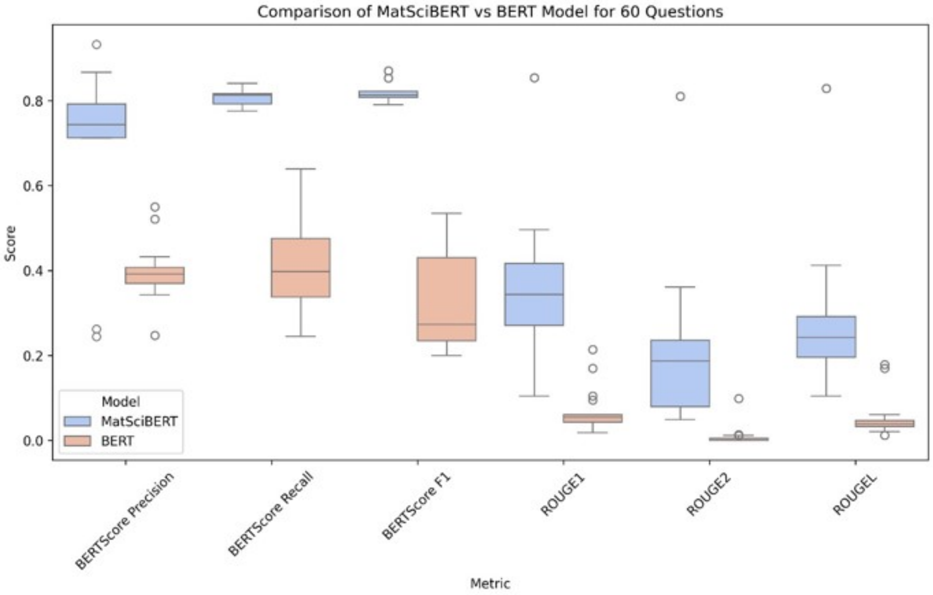
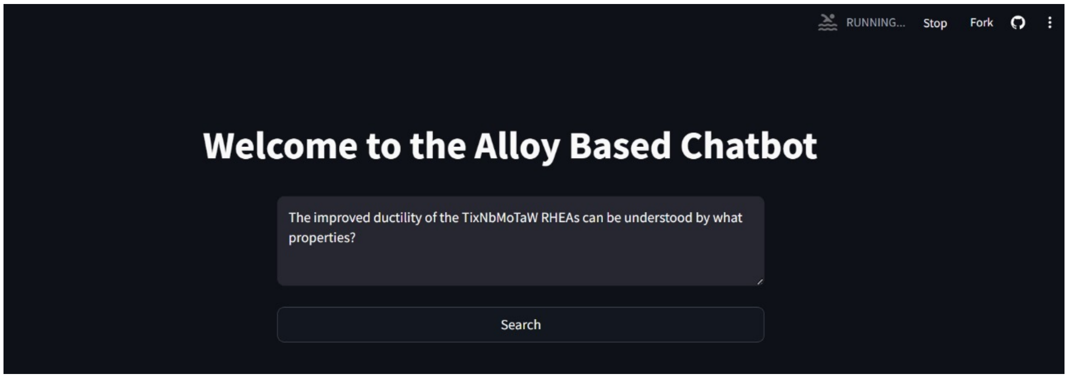


Figure 3.7: Comparison of MatSciBERT and BERT Model for 60 Questions

Our MatSciBERT-based RAG model demonstrated significant improvements over BERT in scientific question answering. It achieved higher BERTScore and ROUGE scores, better alignment with benchmark responses, and improved factual consistency. The combination of optimized retrieval, structured context formatting, and response validation enabled robust, domain-specific question answering without retraining, confirming the effectiveness of our pipeline.

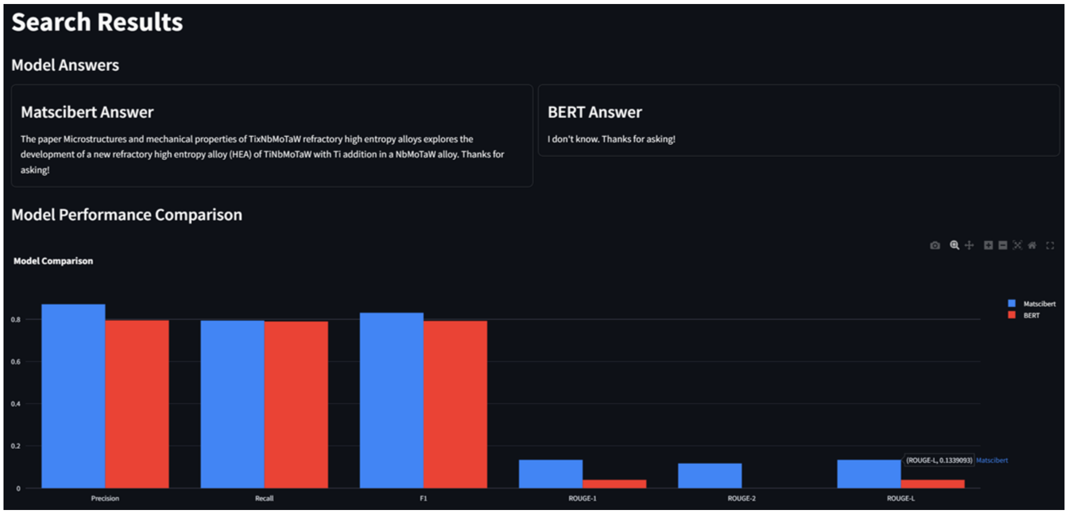
3.5.2 Chatbot in Streamlit



Welcome to the Alloy Based Chatbot

The improved ductility of the TixNbMoTaW RHEAs can be understood by what properties?

Figure 3.8: User Interface



Search Results

Model Answers

Matscibert Answer

The paper Microstructures and mechanical properties of TixNbMoTaW refractory high entropy alloys explores the development of a new refractory high entropy alloy (HEA) of TiNbMoTaW with Ti addition in a NbMoTaW alloy. Thanks for asking!

BERT Answer

I don't know. Thanks for asking!

Model Performance Comparison

Figure 3.9: Generated Answer

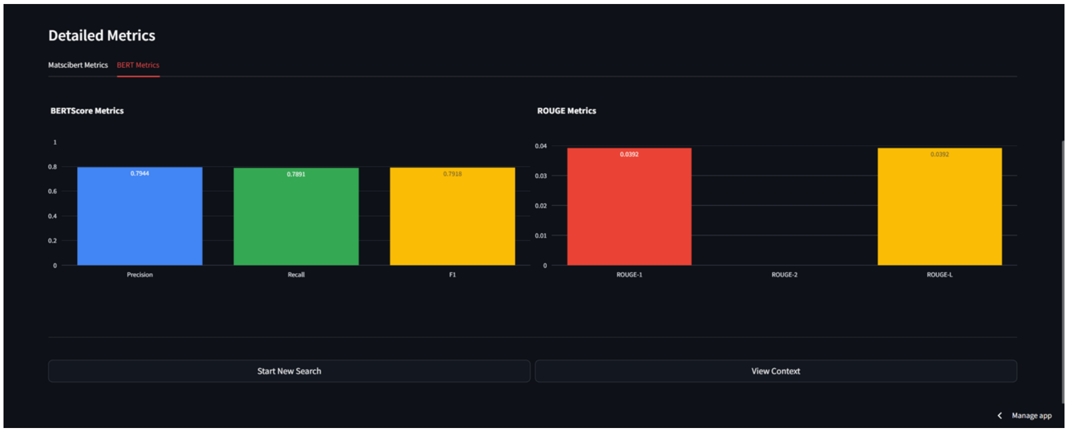


Figure 3.10: BERT Scores

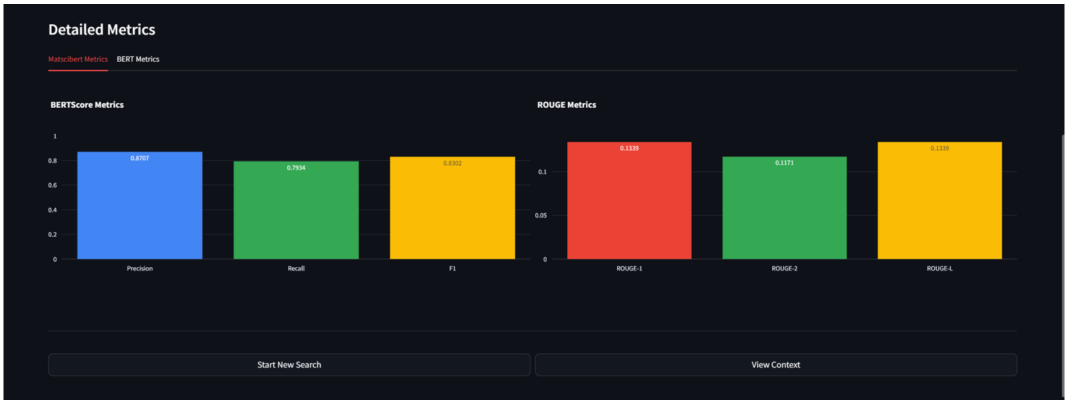
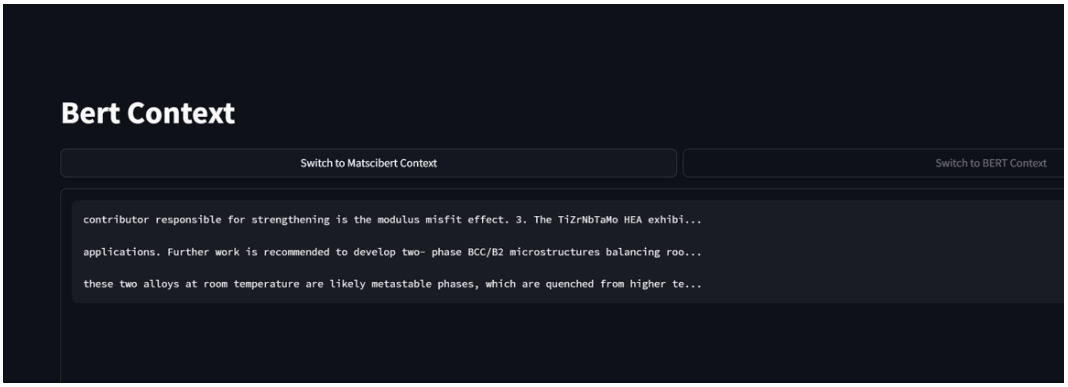


Figure 3.11: MatSciBERT Scores

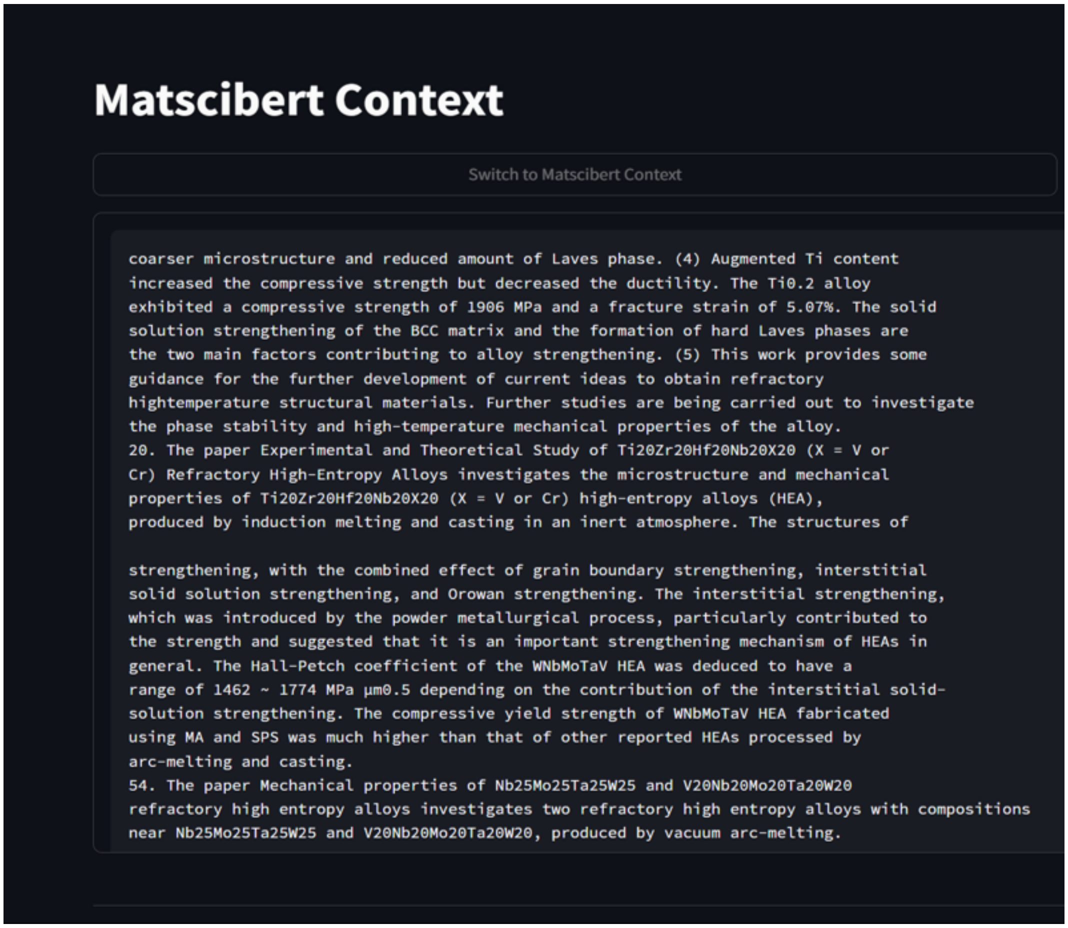


```
contributor responsible for strengthening is the modulus misfit effect. 3. The TiZrNbTaMo HEA exhibi...

applications. Further work is recommended to develop two- phase BCC/B2 microstructures balancing roo...

these two alloys at room temperature are likely metastable phases, which are quenched from higher te...
```

Figure 3.12: BERT Content



```
coarser microstructure and reduced amount of Laves phase. (4) Augmented Ti content
increased the compressive strength but decreased the ductility. The Ti0.2 alloy
exhibited a compressive strength of 1906 MPa and a fracture strain of 5.07%. The solid
solution strengthening of the BCC matrix and the formation of hard Laves phases are
the two main factors contributing to alloy strengthening. (5) This work provides some
guidance for the further development of current ideas to obtain refractory
hightemperature structural materials. Further studies are being carried out to investigate
the phase stability and high-temperature mechanical properties of the alloy.
20. The paper Experimental and Theoretical Study of Ti20Zr20Hf20Nb20X20 (X = V or
Cr) Refractory High-Entropy Alloys investigates the microstructure and mechanical
properties of Ti20Zr20Hf20Nb20X20 (X = V or Cr) high-entropy alloys (HEA),
produced by induction melting and casting in an inert atmosphere. The structures of

strengthening, with the combined effect of grain boundary strengthening, interstitial
solid solution strengthening, and Orowan strengthening. The interstitial strengthening,
which was introduced by the powder metallurgical process, particularly contributed to
the strength and suggested that it is an important strengthening mechanism of HEAs in
general. The Hall-Petch coefficient of the WNbMoTaV HEA was deduced to have a
range of 1462 ~ 1774 MPa µm0.5 depending on the contribution of the interstitial solid-
solution strengthening. The compressive yield strength of WNbMoTaV HEA fabricated
using MA and SPS was much higher than that of other reported HEAs processed by
arc-melting and casting.
54. The paper Mechanical properties of Nb25Mo25Ta25W25 and V20Nb20Mo20Ta20W20
refractory high entropy alloys investigates two refractory high entropy alloys with compositions
near Nb25Mo25Ta25W25 and V20Nb20Mo20Ta20W20, produced by vacuum arc-melting.
```

Figure 3.13: MatSciBERT Content

Chapter 4

Conclusion

The major findings of this work span across two key phases,each tackling severe challenges in the design of High-Entropy Alloy (HEA). During Phase 1 (Design of High Entropy Alloys by using Machine Learning driven Bayesian Optimization), we established an ML-based Bayesian Optimization framework that could effectively survey the enormous compositional space of HEAs efficiently. This methodology enabled us to forecast optimal compositions of alloys with minimal experimental inputs, greatly avoiding costly and labor-intensive trial-and-error approaches. Yet, although this ML-based model successfully captured tabular numerical information like composition, test temperature, and important properties like yield strength to density ratio, it had no ability to include important unstructured information like phase formation, synthesis routes, and characterization processes. To appreciate this constraint, we moved into Phase 2, wherein we built a Retrieval-Augmented Generation (RAG)-informed knowledge extraction system to better facilitate the access and usefulness of scientific literature to HEA design. By including domain-embodied embeddings (MatSciBERT) and sophisticated natural language generation (Llama 3.1), the system allows scientists to access, integrate, and engage with scientific information in a structured fashion. The retrieval process utilizes a vector database (ChromaDB) for semantic search to provide contextually relevant and scientifically sound answers. We extended this system towards ease of use and accessibility through an interface in the form of a chatbot via Streamlit, facilitating easy interaction with the knowledge extraction model. The system's effectiveness in retrieving and producing accurate scientific insights is validated through BERT Score, ROUGE, and Cosine Similarity metrics. By combining machine learning-optimized optimization in Phase 1 and retrieval-based knowledge extraction in Phase 2, our work offers a comprehensive framework for speeding up HEA discovery. Refining retrieval mechanisms, increasing the dataset for better generalization, and optimizing computational efficiency for large-scale deployment will be the focus of future work. This research highlights the promise of AI-based approaches to transform research in materials science, filling the gap between data-driven structured optimization and unstructured knowledge extraction.

References

1. J.-P. Couzinié, O.N. Senkov, D.B. Miracle, G. Dirras, *Comprehensive data compilation on the mechanical properties of refractory high-entropy alloys*, Materials Science and Engineering A 769 (2019) 138527.

2. A. Chandrasekhara, J. Chan, F. Ogoke, O. Ajenifujah, A. B. Farimani, *AMGPT: a Large Language Model for Contextual Querying in Additive Manufacturing*, arXiv preprint arXiv:2406.00031v1 [cs.CL], (2024).

3. T. Gupta, M. Zaki, N. M. A. Krishnan, Mausam, MatSciBERT: *A materials domain language model for text mining and information extraction*, npj Computational Materials 8 (2022) 102.

4. Y. Song, S. Miret, B. Liu, MatSci-NLP: Evaluating Scientific Language Models on Materials Science Language Tasks Using Text-to-Schema Modeling, arXiv:2305.08264v1 [cs.CL] (2023).

5. M. V. Koroteev, BERT: A Review of Applications in Natural Language Processing and Understanding, Financial University under the government of the Russian Federation (2023).

6. Gillioz, J. Casas, E. Mugellini, O. A. Khaled, Overview of the Transformer-based Models for NLP Tasks, Proceedings of the Federated Conference on Computer Science and Information Systems 21 (2020) 179–183.

7. S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, M. Shah, Transformers in Vision: A Survey, ACM Computing Surveys (2022) 0360-0300.

8. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, *Transformers: State-of-the-Art Natural Language Processing*, Proceedings of the 2020 EMNLP (Systems Demonstrations), (2020) 38–45.

9. H. Li, Y. Su, D. Cai, Y. Wang, L. Liu, *A Survey on Retrieval-Augmented Text Generation*, arXiv preprint arXiv:2202.01110v2 [cs.CL], (2022).

10. P. Zhao, H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, L. Yang, W. Zhang, J. Jiang, B. Cui, *Retrieval-Augmented Generation for AI-Generated Content: A Survey*, arXiv preprint arXiv:2402.19473v6 [cs.CV], (2024).

11. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, D. Kiela, *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, arXiv preprint arXiv:2005.11401v4 [cs.CL], (2021).

12. Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, H.

Wang, *Retrieval-Augmented Generation for Large Language Models: A Survey,* arXiv preprint arXiv:2312.10997v5 [cs.CL], (2024).

13. Z. Jiang, F. F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, Y. Yang, J. Callan, G. Neubig, *Active Retrieval-Augmented Generation*, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), (2023) 7969–7992.

14. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, arXiv preprint arXiv:1810.04805v2 [cs.CL], (2019).

List of Publications based on this research work

1. M. Sai Rahul, Ashwin Devan, Kurakula Prashanth, M. Srinivasa Sai Kumar Reddy, and Kritesh Kumar Gupta, *Design of High Entropy Alloys by using Machine Learning driven Bayesian Optimizatio*n, 5th International Conference on Current Trends in Materials Science and Engineering,(Presented).