Automated Data Extraction from Materials Science Literature on High-Entropy Alloys for Efficient Materials Research

PROJECT REPORT

Submitted by

Aman Sirohi - (CB.EN.U4AIE21003)
R Sriviswa - (CB.EN.U4AIE21046)
Souvik Gorain - (CB.EN.U4AIE21065)
Vikhyat Bansal - (CB.EN.U4AIE21076)

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING
(ARTIFICIAL INTELLIGENCE)

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

COMPUTER SCIENCE AND ENGINEERING ARTIFICIAL INTELLIGENCE

AMRITA SCHOOL OF ARTIFICIAL INTELLIGENCE

AMRITA VISHWA VIDYAPEETHAM

COIMBATORE - 641 112 (INDIA)

APRIL - 2025

COMPUTER SCIENCE ENGINEERING - ARTIFICIAL INTELIIGENCE

AMRITA VISHWA VIDYAPEETHAM

COIMBATORE - 641 112

BONAFIDE CERTIFICATE

This is to certify that the thesis entitled **“Automated Data Extraction from Materials Science Literature on High-Entropy Alloys for Efficient Materials Research”** submitted by **Aman Sirohi (CB.EN.U4AIE21003), R Sriviswa (CB.EN.U4AIE21046), Souvik Gorain (CB.EN.U4AIE21065), Vikhyat Bansal (CB.EN.U4AIE21076)** for the award of the **Degree of Bachelor of Technology** in the **“COMPUTER SCIENCE ENGINEERING - ARTIFICIAL INTELIIGENCE”** is a bonafide record of the work carried out by them under my guidance and supervision at Amrita School of Artificial Intelligence, Coimbatore.

Dr Kritesh Gupta
Project Guide(s)
Assistant Professor

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

INTERNAL EXAMINER **EXTERNAL EXAMINER**

AMRITA SCHOOL OF ARTIFICIAL INTELLIGENCE

AMRITA VISHWA VIDYAPEETHAM

COIMBATORE - 641 112

DECLARATION

We, **(CB.EN.U4AIE21003, CB.EN.U4AIE21046, CB.EN.U4AIE21065, CB.EN.U4AIE21076),** hereby declare that this thesis entitled **"Automated Data Extraction from Materials Science Literature on High-Entropy Alloys for Efficient Materials Research"**, is the record of the original work done by us under the guidance of **Dr Kritesh 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 heartfelt thanks to everyone who helped us during this project. First, we would like to thank our institution Amrita Vishwa Vidyapeetham, Coimbatore for providing us with the necessary resources and great environment to do our research in the best way possible. We also acknowledge our project guide, Dr Kritesh Gupta, whose expertise and knowledge encouraged us to simplify our selected topics. We have all been inspired by his knowledge and thoughts in our work. We also thank our faculty and peers for inspiring and engaging discussions that helped shape and inform our thinking on the topic. Additionally, we thank the authors and researchers whose prior work built the groundwork for our analysis to be established. They have played a crucial part in helping us further our research. Finally, we thank our families and friends for their encouragement and support throughout all stages of this work. It has been a pleasant learning experience, and we sincerely thank everyone who contributed to this project.

Team - A1

- Aman Sirohi CB.EN.U4AIE21003
- R Sriviswa CB.EN.U4AIE21046
- Souvik Gorain CB.EN.U4AIE21065
- Vikhyat Bansal CB.EN.U4AIE21076

List of Figures

List of Tables

List of Abbreviations

Serial No.	**Full Form**	**Abbreviation Used**
1	Large Language Model	LLM
2	Natural Language Processing	NLP
3	Retrieval Augmented Generation	RAG
4	Named Entity Recognition	NER
5	True Positive	TP
6	False Positive	FP
7	True Negative	TN
8	False Negative	FN
9	JavaScript Object Notation	JSON
10	Text-To-Text Transfer Transformer	T-5
11	Machine Reading Comprehension	MRC
12	Table-To-Text	ToTTo

Abstract

High-Entropy Alloys (HEAs) are a revolutionary breakthrough in the field of materials science known for their outstanding mechanical properties, thermal stability, and resistance to environmental degradation. Yet, something yet solving an iterative cycle of exploration and optimization found in HEAs due to their vast compositional space and a plethora of numerous relevant research data. The design of HEAs by classical experimental methods is labor and resource-intensive, highlighting the need for AI tools for efficient literature mining and data extraction. This project aims to develop an AI-powered framework for automating the retrieval, annotation, and analysis of HEA-related research literature. By applying state-of-the-art Natural Language Processing (NLP) methodologies, such as Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG), the system methodically collects critical information regarding alloy compositions, structural features, and mechanical properties from scientific literature. This data is extracted and organized in a structured way in a knowledge base to recognize trends, rationalize alloy compositions, and accelerate material discovery time. The study focuses on the performance of various AI methodologies in analyzing domain-specific literature accurately, efficiently, and at a scale. This project seeks to connect state-of-the-art machine learning methods with unstructured scientific literature and mass data-driven HEA research. These results set a precedent for applying AI-augmented literature mining to accelerate materials informatics to the benefit of high-throughput alloy design.

Chapter 1

Introduction

Recently, the development of novel materials, such as High-Entropy Alloys (HEAs), have emerged as a new class of material with improved mechanical and thermal properties. Due to their remarkable strength, ductility, wear and corrosion resistance. These alloy systems are potential candidates for use in aerospace, structural, and energy applications. Nonetheless, the immense compositional space of HEAs poses hurdles in data management, analysis, and optimization. Conventional experimental methods used for HEA discovery and characterization can be costly, resource-demanding and time-consuming. In order to address these issues, researchers are increasingly utilizing artificial intelligence (AI) and machine learning (ML) methods for automated literature-mining, data extraction, and prediction of optimal HEA compositions. With the capability of Natural Language Processing (NLP) and the advantages of GPT-based Large Language Models (LLMs), it is helpful to screen relevant information from various scientific literature and facilitate initial selection of new HEAs with improved properties. The project will develop an AI-based framework to automatically annotate and extract HEA-related information from the scientific literature. With the help of state-of-the-art language models like GPT, Retrieval-Augmented Generation (RAG), and few-shot prompting strategies, the proposed system would consolidate the materials science publications, identify significant alloy attributes and create organized json datasets for additional analysis. Additionally, the research will benchmark various AI approaches as well as determining how accurate, efficient, and scalable each can manage to extract useful information from materials science datasets. We aim to set a foundation for literature analysis of HEAs that is solid both in terms of cost-efficiency and scalability that can serve as a cornerstone in the development of the field of materials informatics. These structured insights will not just help streamline HEA research through the generation of data-driven descriptors for materials design in specialized structural space, they will simultaneously provide an experimental guidance map for future work.

1.1 Literature Survey

- **TSQA: Tabular Scenario-Based Question Answering**

  The paper TSQA: Tabular Scenario-Based Question Answering authored by Xiao Li, Yawei Sun, and Gong Cheng, published in arXiv Computation and Language on January 14, 2021, introduces Tabular Scenario-Based Question Answering (TSQA). Table-based QA includes answering questions that involve extracting and reasoning over information across multiple cells in a table. It also uses text passages and domain knowledge to correct the responses, ensuring the correctness of the responses.

  In order to enable more research in this area, the authors have also created a dataset, called GeoTSQA which in turn specializes in TSQA task for geography domain. A significant aspect of their approach is a new Table-to-Text Generating (TTGen) module that transforms structured table data into readable sentences. Reformulating the Q/A pairs as such facilitates MRC to efficiently extract answers.

  Although it shines a contribution, the study also has several limitations. Higher training and inference cost is a result of increased dataset size and scenario complexity. Furthermore, neither GeoTSQA is covering the whole field of geography, nor this corpus can potentially reflect real criteria of question-answering with a high variability.

- **FeTaQA: Free-Form Table Question Answering**

  Free-Form Table Question Answer Pair Generation (FeTaQa), published in MIT Press Direct on 1 January 2022, is the work of Linyong Nan, Chiachun Hsieh,

Ziming Mao, Xi Victoria Lin, and Neha Verma. They based their methodology on the ToTTo dataset, which contains wikipedia tables with corresponding textual descriptions. In contrast to the traditional fact based question answering, FeTaQA is concerned with generating questions based on querying of structured information, thus increasing complexity.
The paper presents two main modeling frameworks for the task. The Pipeline Model is a 2-step approach where a table semantic parser first selects relevant table cells, then a data-to-text generator generates the answer. The End-to-End Model, by contrast models the task as a sequence-to-sequence task where both a question and table are taken as input to produce free-form answers directly.
But the study does have some limitations. Most existing datasets are dominated by short-answer questions and therefore do not allow for evaluation of more elaborate question answering that requires integration of data and deeper reasoning. Finally, generating questions from table cells with undefined semantic relationships, where automatic models struggle for high quality annotation in datasets, leading to annotation challenges.

- **MatSciBERT: A Materials Domain Language Model for Text Mining and Information Extraction**
  MatSciBERT: A Materials Domain Language Model for Text Mining and Information Extraction, published in npj: Computational Materials on May 3, 2022, is the work of Tanishq Gupta, Mohd Zaki, N. M. Anoop Krishnan, and Mausam. The study introduces specialized language model MatSciBERT and adopts a process-based research paradigm, especially that of corpus construction and pre-training as well as downstream tasks.
  In terms of corpus development, the authors created the Materials Science Corpus (MSC), which consists of relevant research papers that provide a domain-specific text dataset. For the pre-training phase, MatsciBERT was trained using MSC as a learning set to expose the model to materials science specific terminology, relations and concepts. Then the model was trained on to perform several downstream tasks such as NER, Abstract Classification, and Relation Classification to

showcase its abilities to handle, interpret, and extract crucial information from the scientific research papers.

While enumerating its contributions, the study however points out some limitations. The approach used presents a critical challenge because, in general, the amount of available annotated datasets in materials science is limited, which inhibits the fine-tuning and evaluation of specialized domain tasks for the model. Furthermore, MatSciBERT is also dependent on SciBERT's vocabulary, which is not capable of fully covering the wide spectrum of terms, symbols, and notations used in materials science publications.

- **A General-Purpose Material Property Data Extraction Pipeline from Large Polymer Corpora Using Natural Language Processing**

  An NLP-based automated pipeline for the extraction of material property information from a large corpus of abstracts was reported in npj Computational Materials (29 September 2022) by Pranav Shetty, Arunkumar Chitteth Rajan, ChristopherKuenneth,Sonkakshi Gupta, L. P. Panchumarti, Lauren Holm, Chaoran Zhang, Rampi Ramprasad. For material discovery, this paper focuses on pre-training a domain specific language model, implementation of Named Entity Recognition (NER) and structuring the data obtained. Central of this work is MaterialsBERT, a domain-specific language model pretrained on a vast material science literature.

  After extracting the material property data it will undergo heuristic-based normalization and structuring, which greatly contributes to establishing a structured material property database. This database facilitates data collection and improves material screening and discovery tasks.

  While it is relatively effective, the study does have a crucial drawback in the polymer name normalization part. This method is based on a static lookup table with a set of common polymer names, which means that it cannot identify and standardize less-known polymer names or abbreviations that does not appear often in the literature. This results in possible limitations on the extraction of the data, as polymer names may not be recognized by model.

- **AMGPT: A Large Language Model for Contextual Querying in Additive Manufacturing**

  This paper released on arXiv by Achuth Chandrasekhar, Jonathan Chan, Francis Ogoke, Olabode Ajenifujah, and Amir Barati Farimani on May 24th, 2024, presents AMGPT, a Large Language Model (LLM) that is both tailored to answer metal Additive Manufacturing (AM) questions yet pre-trained on general knowledge. Our approach is based on advanced language modeling methods proposed in the literature and directly applicable to the AM domain as well.

  AMGPT is based on Llama-2 7B with Retrieval Augmented Generation (RAG): using knowledge sources from databases and the internet to answer context-specific questions. We trained the model on 50 papers and books devoted specifically to additive manufacturing, giving it a domain-aware knowledge base. Using this method, AMGPT can optimize information retrieval to create tighter responses to AM-based inquiries.

  We have noted some limitation of this study. Firstly the no. of research articles utilized for training is very limited. This may affect model performance. Secondly the dependence on highly cited sources over new publications may create certain bias which can affect model's ability to include latest insights.

1.2 Problem Statement

To define and optimize High-Entropy Alloys (HEAs), there exists a massive compositional space and complex mechanical and thermal properties, which make finding them a non-trivial challenge even after 30 years of research. Conventional experimental methods for HEA investigation are profoundly resource-consuming, as they necessitate extensive laboratory testing and computational simulations in order to find optimal alloy compositions. Not only is this process time-consuming, but it also restricts the ability to rapidly explore a wide variety of material configurations. One important challenge in HEA research is the **lack of structured & centralized data** source that captures properties, compositions, and performance metrics of the classes of materials. The vast majority of knowledge available is spread out over thousands of scientific papers, patents and reports, making it hard for researchers to systematically sift through existing research. Considering a collection of state-of-the-art HEA literature, there is currently no standardized methodology to **extract, structure, and mine** HEA-related literature that ultimately hinders data-driven design of alloys. Additionally, the **domain-specific** language and **contextual intricacies** in materials science literature are not well-handled by standard text-mining techniques. As results from current AI and NLP models had shown potential in parsing through this large amount of textual data, the effectiveness of such models in efficiently extracting HEA-specific information in complete accuracy still needs to be explored further. To address that, with no reliable automated literature mining system, researchers struggle to follow trends, perform HEA composition-property search correlation, and expedite material discovery.

1.3 Objectives

This project's main goal is to create an artificial intelligence framework for automatic literature mining and knowledge extraction about High-Entropy Alloys (HEAs). This project addresses the need to overcome the vast compositional space that exists for HEAs, the unstructured nature of research data, as well as inefficiencies in traditional discovery methods. This project mainly aims at the following goals:

- **Automated Literature Retrieval and Analysis**
  - To create an NLP-based system for high-throughput data mining and analysis of the scientific literature relevant to HEAs.
  - Use Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) for extracting useful information from abstracts.
- **Building the Data Pipelines — Extracting and Structuring the Data**
  - Data mining of abstracts from published research for alloy compositions, mechanical properties, structural characteristics, performance metrics.
  - Put the extracted data into an organized, searchable format to enable further analytics.

- **A Knowledge Base for HEA Research**
  - Centralize extracted HEA information in a structured format to facilitate data-driven materials discovery.
  - Enable linking between alloy composition and mechanical/thermal performance via pattern recognition.
- **Improving the Efficiency of HEA Research and Development**
  - Reduce time and effort associated with manually reviewing literature associated with HEA research.
  - Incorporating actionable insights from scientific literature that provide more metrics and less uncertainty.
- **Evaluate Performance:**
  - A comparison will be made to assess whether compared to existing norms if AI can make literature review more effective and efficient in the end.
  - Confidence that the proposed systems can be scaled and used in future materials informatics work.
- **Development of an interactive Chat Bot**
  We aim to create a chatbot which allows users to perform the following tasks:
  - Input abstracts both as a csv [multiple] and text [single].
  - Display material science based properties of HEA alloys in tabular form [downloadable JSON].
  - Generate suggestive Question-Answer pairs from the abstracts given as input.
  - In the meantime as Question-Answer pairs are getting generated users should be able to interact with data and ask custom-made questions about HEA alloys to our chatbot.

Chapter 2

Background

2.1 High Entropy Alloys (HEA)

High-Entropy Alloys (HEAs) are a new class of metallic materials with remarkable mechanical properties, thermal stability, and corrosion resistance. In contrast to traditional alloys, which draw on one or two principal elements, HEAs are composed of an ensemble of elements of nearly equally proportioned atomic concentrations. This composition results in unique microstructure characteristics such as solid-solution strengthening, lattice distortion and stability of phases, which makes HEAs suitable for a range of applications including aerospace, energy and structural engineering.

2.2 Challenges in HEA Discovery and Optimization

Nevertheless, the discovery and optimization of HEAs are challenging due to the immense composition space and complex alloy-property and structural behavior relationships. Traditional methodologies for discovering advanced HEAs involve time-consuming synthesis and testing associated with developing new materials, leading to inefficiency and limited extended-exploration capacity. Computational tools, specifically thermodynamic modeling and first-principles calculations, have also enhanced understanding, but such tools are limited by availability of structured experimental data.

2.3 Fragmented & Unstructured Knowledge and the need for Insight Extraction and Organization

A key problem in HEA research is the absence of a centralized, structured knowledge base. Information about HEAs is distributed across many research articles, patents, and technical reports, making it hard to access insights directly. Often, manual literature reviews take a long time with no guarantee of relevant material insights, and a basic search of database studies fails to identify precise material properties, because much of the research studies, tips and reviews use field-specific and contextual descriptors. Thus, insight extraction methods are needed to enable the systematic review and organization of HEA information in a manner that is conducive to rapid discovery of materials.

2.4 Role of AI and NLP in HEA Research

The emergence of new technologies in (AI) and (NLP) promises to enable automation of information extraction from scientific publications, introducing entirely new pathways for accelerating research in HEA. AI-based literature mining could dramatically improve the productivity of HEA research by facilitating the identification of trends and allowing the association of alloy composition with mechanical property data to enable data-driven material design. Using (LLMs) and (RAG) researchers will also be able to process larger and larger amounts of unstructured texts, extract important points of data and build comprehensive databases to evaluate the HEA literature.

2.5 AI Automated Framework for HEA Research

This proposal will address these challenges so that it sets out to build an AI Automation Framework for literature retrieval, annotation, and knowledge extraction as applied to HEAs. The proposed AI-based system will systematically identify and analyze scientific research papers and papers in order to extract alloy composition, mechanical properties, and structural parameters for systematic structuring into a searchable knowledge base.

Chapter 3

Methodology

High-Entropy Alloys (HEAs) have attracted a lot of attention thanks to their excellent mechanical properties, corrosion resistance, and high-temperature performance. Uncovering and optimizing HEAs is still a difficult endeavor due to the vast compositional space of HEAs and the absence of general techniques used to generate structured knowledge. Traditional experimental approaches can also be lengthy and expensive from a computational perspective, highlighting the need for automated approaches to mine the literature and perform data-driven investigations.

This project seeks to alleviate this challenge by creating an AI-facilitated literature mining framework to systematically extract and organize HEA-related data from scientific publications. The proposed effort extends work that has already been advanced in the area by using Artificial Intelligence to help improve techniques of extracting and structuring knowledge.

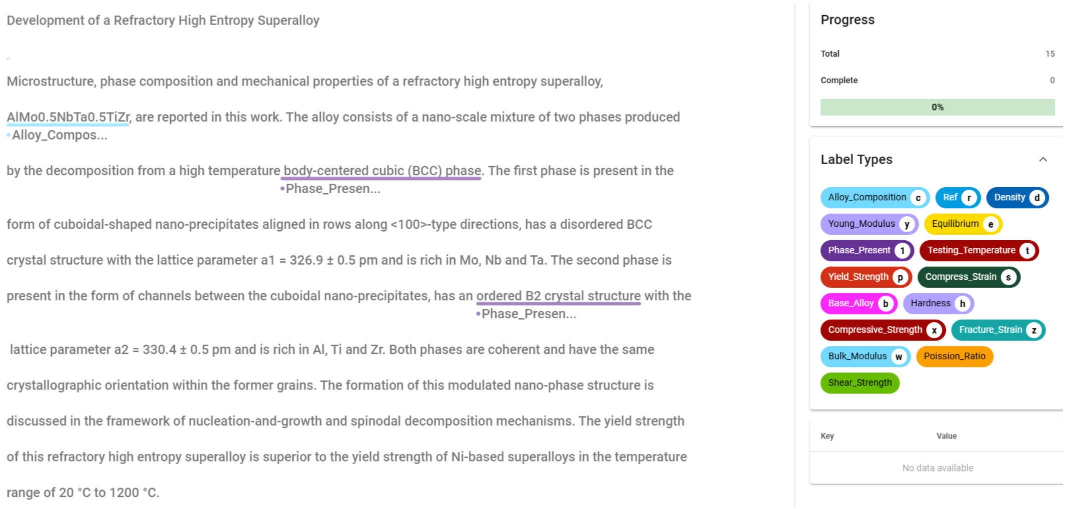
3.1 Stage-1

3.1.1 Knowledge Extraction in JSON Format

The first stage of this research revolves around the extraction, pre-processing, and organization of relevant data from sources in a systematic way. This stage is designed to make sure that the extracted data is clean, organized and ready for subsequent processing in the QA system. The primary steps of this stage are:

3.1.1.1 Data Collection

- Identify research manuscripts and documents that present some necessary data. The documents are then sourced from a community-powered and relevant scientific repository.
- Documents that are collected are then processed into machine-readable formats to allow for automated processing.



Development of a Refractory High Entropy Superalloy

Microstructure, phase composition and mechanical properties of a refractory high entropy superalloy, AlMo0.5NbTa0.5TiZr, are reported in this work. The alloy consists of a nano-scale mixture of two phases produced by the decomposition from a high temperature body-centered cubic (BCC) phase. The first phase is present in the form of cuboidal-shaped nano-precipitates aligned in rows along <100>-type directions, has a disordered BCC crystal structure with the lattice parameter a1 = 326.9 ± 0.5 pm and is rich in Mo, Nb and Ta. The second phase is present in the form of channels between the cuboidal nano-precipitates, has an ordered B2 crystal structure with the lattice parameter a2 = 330.4 ± 0.5 pm and is rich in Al, Ti and Zr. Both phases are coherent and have the same crystallographic orientation within the former grains. The formation of this modulated nano-phase structure is discussed in the framework of nucleation-and-growth and spinodal decomposition mechanisms. The yield strength of this refractory high entropy superalloy is superior to the yield strength of Ni-based superalloys in the temperature range of 20 °C to 1200 °C.

Progress

Total	15
Complete	0

0%

Label Types

Key | Value

No data available

Figure 3.1: Illustration showing the annotated text as well as labels using doccano.

3.1.1.2 Knowledge Extraction from Data

We present a research paradigm that is constructed to swiftly analyze the scientific abstracts from the Elsevier Scopus database that are specifically directed towards all pertinent scientific knowledge that involves high-entropy alloys (HEAs). It combines various approaches to provide a robust extraction of domain data that also converts them into structured JSON format. The process starts with data collection and preprocessing, where they retrieve the relevant abstracts using the filtering features offered by Elsevier. These abstracts are then cleaned by purging irrelevant metadata, standardizing formats, and applying basic preprocessing steps such as converting all to lowercase, punctuation removal and tokenization. We perform manual annotation using the Doccano tool with the help of a materials science expert to create a high-quality labeled dataset which is essential for the fine-tuning of the LLM and serves as a gold standard for evaluation.

The first approach utilizes LLaMA 3.1 8B from the Hugging Face library that is trained on the annotated dataset to learn domain-specific nuances in HEA literature. This fine-tuned model analyzes abstracts to identify elements, chemical compositions, alloy designations and main conclusions and to classify abstracts into research focus. Using the manually annotated abstracts as ground truth, accuracy and F1 scores to evaluate their performance are calculated with respect to match rate for alloy data such as names, compositions, and performance properties.

The second approach uses a RAG (Retrieval-Augmented Generation) method that combines the LLM-based extraction with context confirmation. In this we create dense vector representations for a selected knowledge base of review papers and landmark studies, embedding them into a FAISS vector database, and semantic retrieval for relevant context. Finally, a model is employed to produce structured outputs from the content that was retrieved, allowing for semantic similarity refinement.

In addition, a few-shot prompting strategy is used applying a GPT model. This entails designing focused prompts with a small number of examples to extract information from HEA abstracts. This structured data is extracted with prompts as well as relevant context sent along with it, and then the returned responses from the model are processed and structured to extract data on composition, synthesis methods, properties, and findings specific to alloys. The output is transformed and parsed into standardized JSON, allowing easy integration and the performance of large scale, automated HEA literature analysis.

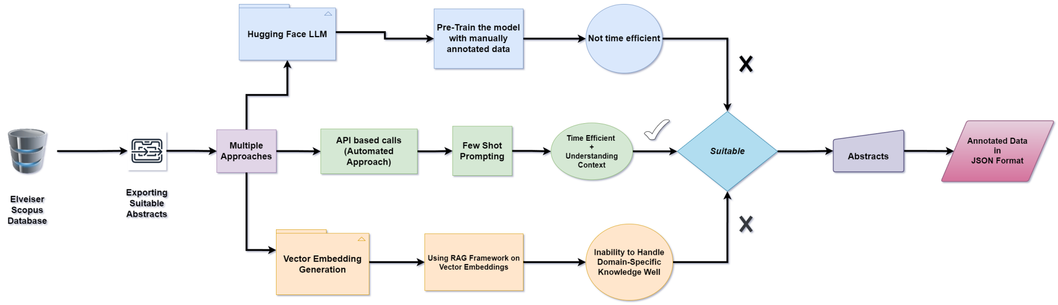

Hugging Face LLM
Pre-Train the model with manually annotated data
Not time efficient
X
Elveiser Scopus Database
Exporting Suitable Abstracts
Multiple Approaches
API based calls (Automated Approach)
Few Shot Prompting
Time Efficient + Understanding Context
✓
Suitable
Abstracts
Annotated Data in JSON Format
X
Vector Embedding Generation
Using RAG Framework on Vector Embeddings
Inability to Handle Domain-Specific Knowledge Well

Figure 3.2: Illustration of proposed methodology.

3.1.1.3 Results from Knowledge Extraction

In order to assess the accuracy of data that is extracted we utilized a simplified and refined confusion matrix based framework. In this frame work we have gold standard annotation that will serve as actual prediction and then we have the output generated by models which will serve as predictions. In our approach the True Positive (TP) will indicate an exact match between the gold standard and received model response. True Negative (TN) will refer to absence of material property in both gold standard and output predicted by model. False Positive (FP) will refer to those values which are generated by model but not present in our golden standard and False Negative (FN) will refer to information that is missing in output generated by model but present in gold standard. By using these we calculated the following performance metrics:

$$\begin{aligned}
\text{Precision} &= \frac{TP}{TP + FP} \\
\text{Recall} &= \frac{TP}{TP + FN} \\
F1\ \text{Score} &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\end{aligned}$$

To evaluate performance across models, five additional examples were generated and compared across models against gold standard annotations generated by domain experts. To deal with the complexity of structured JSON outputs, which each captured varied elements of the scientific elements produced a wrapper function was used to process the outputs by post-processing to return only the required segments that could be evaluated.

Table 3.1: Comparison of Different Language Models

Model Used	Effective Cost Per Output	Output Time	Accuracy	F1 Score
Llama-3.1 8B Parameter	Free	10 mins	56%	0.47
OpenAI GPT-4	0.5	3 mins	91%	0.92
OpenAI GPT-4o	0.15	4 mins	90%	0.88
OpenAI GPT-4o Mini	0.005	1 mins	92%	0.89
OpenAI GPT-o1 Mini	2.5	6 min	85%	0.87
Gemini Pro	0.075	3 min	84%	0.88
Deepseek V3	0.05	2 min	89%	0.85
Claude Sonnet 3.5	4.75	2 min	94%	0.92

3.2 Stage-2

This project describes the construction of an automatic system for producing organized Question & Answer (Q/A) pairs from high-entropy alloy (HEA) datasets in JSON format. To achieve this, we utilize the pretrained large language model (LLM) Qwen2.5-1.5B-Instruct by Alibaba Cloud. The Q/A pairs we produce are then organized based on diverse dimensions related to the properties of alloys, so they are relevant and comprehensive.

3.2.1 Suggesting Question-Answer Pairs:

3.2.1.1 Dataset Preparation and Context Structuring:

- The dataset used in the process is comprised of structured JSON files containing domain knowledge, predominately concentrating on alloys and their properties.
- Contexts are extracted from the JSON file dynamically and contextually negate multiple alloy entries to a single context to allow contextual integrity and process cropping.
- Each alloy's attributes (i.e., composition, performance properties, material conditions, metadata (i.e., DOI)), are organized hierarchically for efficient parsing and retrieval purposes.

3.2.1.2 Model initialization and Tokenization:

- The pretrained language model is initialized with automatic detection of type and device assignment to improve performance.
- A tokenizer is then used to process the input and output sequences to allow the model to capture the information presented in constructing the output.
- Finally, short instruction formats meant to generate quality interactions are used to form the input prompt to ensure structured outputs that are relevant to the content that model produced.

3.2.1.3 Automated Generation of Question Answer pairs:

- A structured prompt is carefully crafted to guide the model to produce varied Q-A pairs for each context that fall into lines:
    - Property Value-Based Questions: Extract numerical values associated with properties of materials.
    - Temperature-Based Questions: Ask questions about the environment in which the properties were measured.
    - Unit-based Question: Identify units of measurement related to the reported properties of the material.
    - Material Condition-Based Questions: Investigate equilibrium state, types of phases and subsequent conditions related to alloys.
- The Qwen model processes the structured prompt and generates responses with a strict defined JSON format to ensure consistency.
- The generated output is immediately processed. Heuristic validation is used to find inconsistencies or deviation from the format.

3.2.1.4 Heuristic and LLM Based Filtering:

- The initial filter is heuristic based to remove unstructured, redundant, or placeholders response.
- The validated Q-A pairs are sent through the same LLM for an assessment where the pairs are rated on criteria for categorical alignment, contextual accuracy, and overall relevance.
- In order to illustrate the effectiveness of filtering, five non-filtered Q-A pairs and the top five rated Q-A pair were randomly shuffled, and human experts were asked to evaluate without knowledge of filtering.

- A performance gain of about 33% in selecting Q-A pairs was observed in the filtering procedure, indicating the filtering and model rating adds to the relevance and contextual validity of the generated Q-A pairs.

3.2.1.5 Inference and Comparison Evaluation:

- The only Q-A pairs grouped into a category that were kept in the final dataset were the pairs of the highest rank. This guarantees the final dataset has high quality, valid entries that fit the context.
- The final output is stored in a standardized JSON format, ensuring seamless integration into downstream applications such as knowledge-driven chatbots, alloy property recommendation systems, and material science databases.

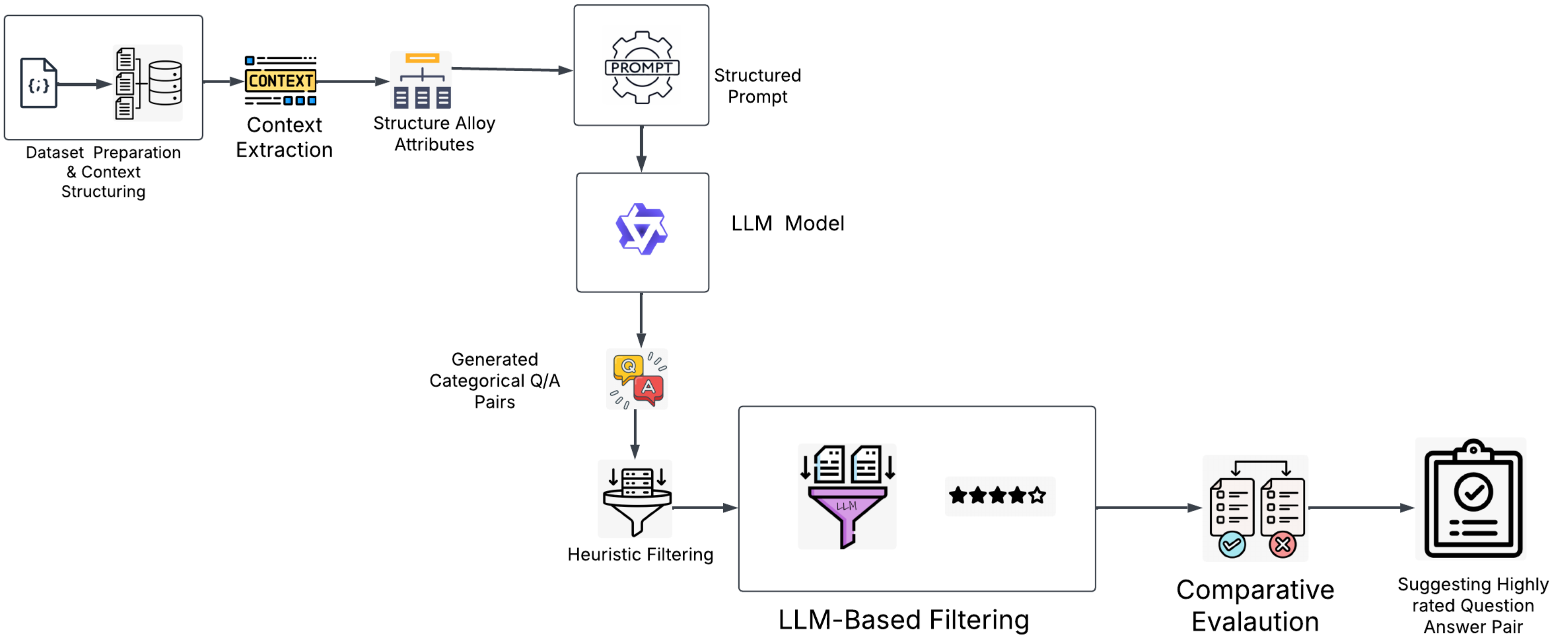

Dataset Preparation & Context Structuring
CONTEXT
Context Extraction
Structure Alloy Attributes
PROMPT
Structured Prompt
LLM Model
Generated Categorical Q/A Pairs
Q
A
Heuristic Filtering
LLM
LLM-Based Filtering
Comparative Evalaution
Suggesting Highly rated Question Answer Pair

Figure 3.3: Pipeline demonstrating Question Answer pair generation.

3.2.2 Model Finetuning:

In addition to generating suggestive QnA pairs, we leveraged the structured data from Stage-1 to finetune several language models specifically for the HEA domain. This finetuning process aims to enhance model performance on domain-specific question answering tasks by adapting pre-trained language models to the specialized vocabulary and knowledge patterns found in high-entropy alloy literature.

3.2.2.1 Dataset Preparation for Model Finetuning:

- The JSON output from Stage-1 was transformed into a specialized dataset (**HEA-QA-DATASET.json**) following the **SQuAD** format, which is the standard for question-answering tasks.
- Each entry in the dataset consists of a context paragraph about specific alloys, associated questions, and the corresponding answer spans within the context.
- This structured format enables the models to learn the relationship between questions and where their answers can be found within the alloy-specific contexts.
- The dataset maintains the technical terminology and specialized knowledge patterns present in materials science literature, ensuring models develop domain-specific understanding.

3.2.2.2 T5-Base Model Finetuning:

- **Model Architecture:** We employed the T5-Base model, a sequence-to-sequence transformer architecture specifically designed for text-to-text tasks, making it well-suited for question answering applications.
- **Implementation Framework:** The finetuning was implemented using PyTorch Lightning, which provides a structured and scalable approach to managing the training process.
- **Technical Specifications:**
  - Model Size: 223 million parameters

  - Architecture: 12 encoder layers + 12 decoder layers
  - GPU Configuration: NVIDIA T4
  - VRAM Utilization: 16GB

- **Training Parameters:**
  - Epochs: 2
  - Batch Size: 2
  - Learning Rate: 0.0001
  - Optimizer: AdamW
  - Training Time per Epoch: Approximately 20 minutes
  - Total Training Duration: 2 hours

- The sequence-to-sequence approach of T5 allows the model to generate complete answer spans rather than simply identifying positions within the source text.

3.2.2.3 RoBERTa-Base-SQuAD2 Model Finetuning:

- **Model Architecture:** We utilized RoBERTa, a robustly optimized BERT variant, specifically targeting the SQuAD2 version that has shown strong performance on question answering benchmarks.

- **Implementation Framework:** Finetuning was conducted using Hugging Face Transformers and PyTorch, enabling full model parameter updates throughout the network.

- **Technical Specifications:**
  - Model Size: 124 million parameters
  - Architecture: 12 transformer layers
  - GPU Configuration: NVIDIA T4
  - VRAM Utilization: 16GB

- **Training Parameters:**
  - Epochs: 2
  - Batch Size: 8
  - Learning Rate: 0.0001
  - Optimizer: AdamW
  - Training Time per Epoch: 10-12 minutes
  - Total Training Duration: Approximately 2 hours
- The RoBERTa model employs a span-prediction approach that identifies the start and end positions of answer spans within the context, making it particularly effective for extractive question answering tasks.

3.2.2.4 Phi-2 Model Finetuning with QLoRA:

- **Model Architecture:** We utilized Microsoft's Phi-2, a 2.7 billion parameter language model, to explore the capabilities of larger language models in specialized scientific domains like materials science.
- **Resource Constraints and Solutions:** Due to the substantial computational requirements of Phi-2, a complete model finetuning was impractical with available resources. Instead, we implemented QLoRA (Quantized Low-Rank Adaptation), a parameter-efficient finetuning method.
- **Quantization Implementation:**
  - The base model was quantized to 4-bit precision using BitsAndBytes
  - We employed NF4 (normalized float 4) quantization type for optimal precision
  - Double quantization was enabled to further reduce memory requirements
  - Computation was performed in FP16 (half precision) for efficiency
- **LoRA Configuration:**

  - Rank (r): 8 - defining the dimension of low-rank update matrices
  - Alpha: 16 - scaling factor for LoRA updates
  - Target modules: attention matrices only (”q_proj”, ”k_proj”, ”v_proj”, ”o_proj”)
  - LoRA dropout: 0.1 for regularization
  - Only 0.1% of the original model parameters were trained, significantly reducing memory requirements

- **Training Details:**
  - Input format: Context followed by question, with the model trained to generate the answer
  - Batch size: 4 with gradient accumulation
  - Learning rate: 2e-4 with weight decay of 0.001
  - Optimizer: 8-bit AdamW with paged optimization
  - Training utilized gradient checkpointing for additional memory efficiency

- **Advantages of our QLoRA Implementation:**
  - Reduced VRAM usage from over 10GB to under 6GB, enabling training on consumer GPUs
  - Preserved model quality while focusing training on task-relevant parameters
  - Maintained the pre-trained model's general knowledge while adapting to domain-specific HEA terminology
  - Significantly faster training compared to full model finetuning
  - Portable adapters that can be applied to the base model without distributing the full weights

- This approach demonstrates a practical pathway for adapting state-of-the-art language models to specialized scientific domains even with limited computational resources.

3.2.2.5 Comparative Analysis of Finetuned Models:

- Each finetuned model offers distinct advantages in the HEA domain question answering task:
  - **T5-Base** excels at generating complete, coherent answer formulations
  - **RoBERTa-Base-SQuAD2** demonstrates precision in extracting exact answer spans from technical contexts
  - **Phi-2** with QLoRA leverages larger model capabilities while remaining computationally feasible
- The diversity in model architectures and finetuning approaches provides complementary strengths that can be leveraged in ensemble methods or task-specific deployments.
- The finetuned models serve as specialized tools for automatically extracting and validating information from alloy literature, complementing the suggestive QA pair generation process described in **Section 3.2.1**.

Table 3.2: Performance Metrics Comparison

Evaluation Metric	**T-5**	**RoBERTa**	**Phi-2**
Exact Match	0.921	0.801	0.968
F1 score	0.910	0.836	0.946
BLEU score	0.243	0.171	0.342
ROUGE-L	0.862	0.810	0.953

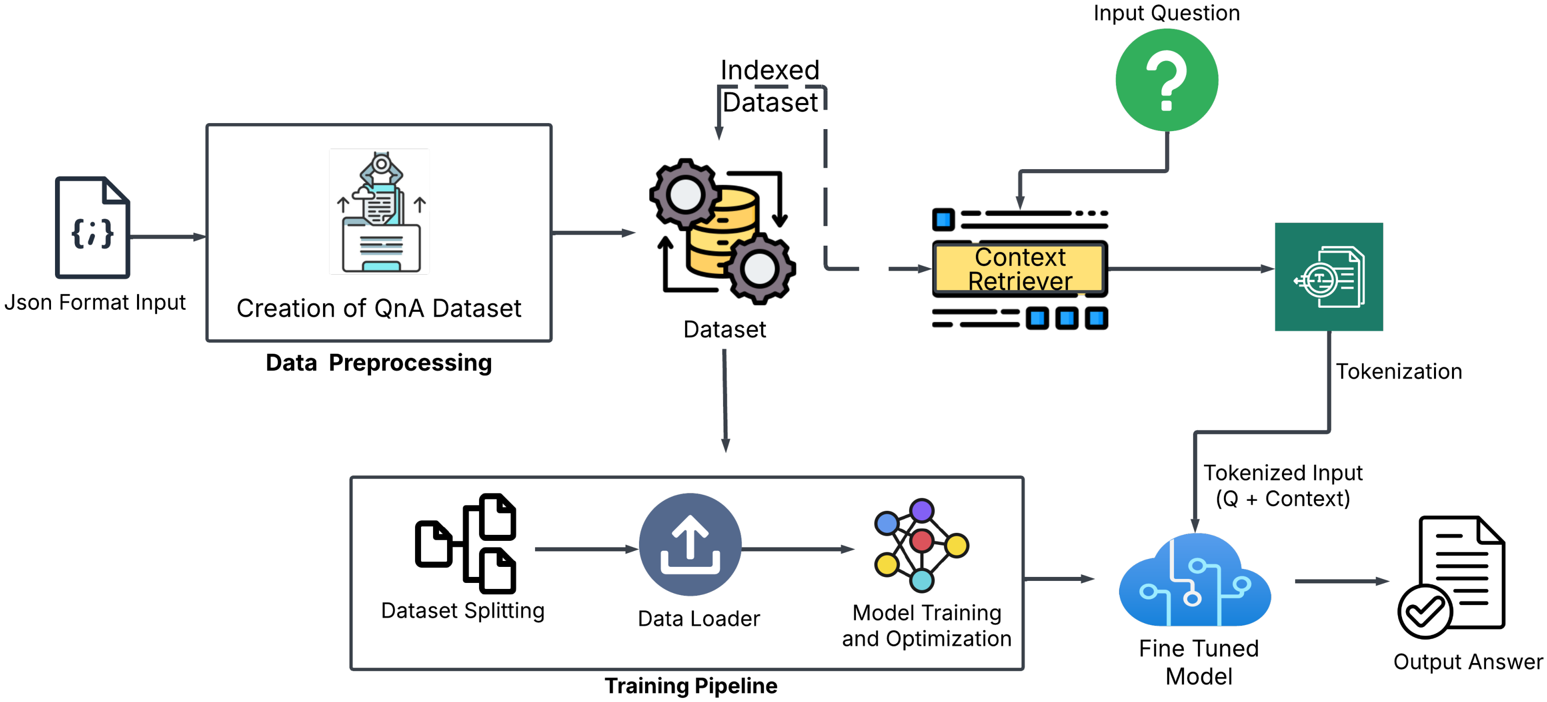


Figure 3.4: Pipeline demonstrating Fine-Tuning and RAG

3.3 Development of Interactive HEA Data Exploration System

Our project culminated in the creation of a web-based application that allows materials scientists to interact with High Entropy Alloy research through an intuitive interface. We've integrated our various project components into what we believe is a practical tool for knowledge discovery in this expanding field.

3.3.1 System Architecture and Implementation

- **Backend Framework:** The application is built using Flask, a lightweight Python web framework, with Flask-Session for persistent state management across user interactions (as it proved more reliable than the alternatives we tested).
- **Component Integration:** The system we developed connects three main capabilities we've worked on throughout this project:
  - Scientific abstract parsing and structured data extraction
  - Automated question-answer pair generation
  - Retrieval-augmented question answering using fine-tuned models
- **Model Deployment:** The application implements a dual-model approach:
  - Local inference using our fine-tuned Phi-2 model with QLoRA adapters
  - API-based inference (a fallback option) for high-performance responses when local resources are limited
- **Memory Management:** Memory management was particularly troublesome during development. We eventually implemented several techniques including document chunking with carefully tuned parameters, quantized model loading to reduce memory footprint, and strategic garbage collection calls that significantly improved stability on our test machines.

3.3.2 Data Ingestion and Processing Pipeline

- **Multiple Input Formats:** Users can submit HEA research data through:
  - Individual abstract submission via text input
  - Batch processing via CSV files containing multiple abstracts
- **Structured Data Extraction:** Submitted abstracts undergo a multi-stage processing pipeline:
  - Initial parsing to identify alloy compositions, properties, and experimental conditions
  - Conversion to standardized JSON format following materials science conventions
  - Organization into tables for visual inspection and analysis
- **Text Chunking for RAG:** The system implements an intelligent text chunking algorithm that:
  - The text chunking algorithm for our retrieval system went through several revisions. Initially, we used fixed-length chunking, but this often split important information across chunks.
  - Our current implementation tries to be smarter - it looks for natural text boundaries such as sentence ends and paragraph breaks, maintains chunks around 384 tokens (which we found optimal after experimentation), and uses a 64-token overlap to prevent information loss at boundaries.

3.3.3 Retrieval-Augmented Generation Implementation

- **Vector Database Integration:** The application implements an in-memory FAISS index rather than alternatives like Annoy or Elasticsearch. This choice was partly pragmatic - FAISS offered the best performance for in-memory operation in our testing, though we recognize that for larger datasets, a persistent storage solution might be necessary. It offers:

  - Fast similarity search across document chunks
  - Session-specific vector stores to maintain context isolation
  - Dynamic re-ranking of retrieved chunks based on relevance scores

- **Context Retrieval Optimization:** The retrieval system incorporates:
  - Adaptive similarity thresholds that adjust based on result quality
  - Duplicate removal to ensure diverse context representation
  - Relevance scoring using cosine similarity between query and context embeddings

- **Answer Generation:** Questions are answered through a specialized prompt structure that:
  - Provides relevant context chunks with clear delineation
  - Instructs the model to synthesize information across multiple sources
  - Directs the model to acknowledge conflicts or missing information

3.3.4 User Interface and Interaction Design

- **Multi-Modal Data Visualization:** The interface presents extracted information through:
  - Interactive, sortable tabular displays of alloy properties and conditions
  - Categorized presentation of generated question-answer pairs
  - Contextual highlighting of information sources for transparency

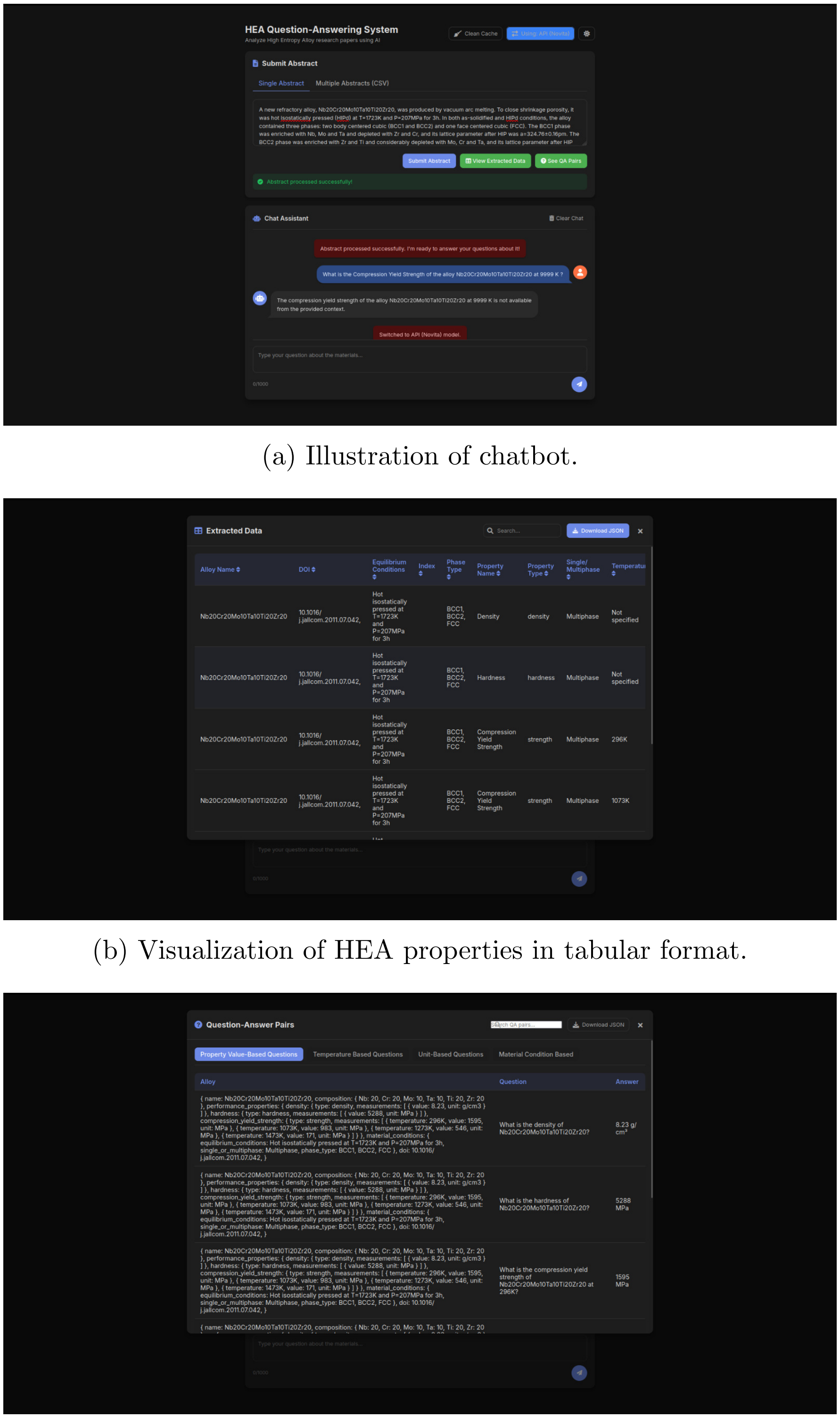
- **Data Export Capabilities:** Users can download processed data in:
  - Structured JSON format for programmatic analysis or database integration
  - Q/A pairs JSON for educational resource development or knowledge base construction

- **Asynchronous Processing:** Resource-intensive operations like Q/A pair generation run asynchronously with:
  - Status indicators for active processes
  - Background processing to maintain UI responsiveness
  - Automatic notification when results are ready for viewing
- **Accessibility Features:** The interface includes:
  - Dark/light mode toggle for different lighting environments
  - Searchable Q/A collections for rapid information retrieval
  - Responsive design for use across different devices

3.3.5 Advanced Features and Technical Optimizations

- **Model Switching Capability:** The system allows seamless switching between:
  - Local fine-tuned Phi-2 model running with quantized weights
  - External API service for compute-intensive queries
- **Session Management:** User sessions incorporate:
  - Persistent storage of processed abstracts and generated Q/A pairs
  - Automated cleanup of old sessions to manage server resources
  - Secure session handling with unique identifiers
- **Error Handling:** The application implements robust error management:
  - Graceful degradation when model loading fails
  - Informative error messages for users
  - Detailed logging for system maintenance and debugging

This integrated system demonstrates how advanced NLP techniques can be applied to create practical tools for materials science research, enabling faster literature analysis and knowledge discovery in the growing field of High Entropy Alloys.



A new refractory alloy, Nb20Cr20Mo10Ta10Ti20Zr20, was produced by vacuum arc melting. To close shrinkage porosity, it was hot isostatically pressed (HIPd) at T=1723K and P=207MPa for 3h. In both as-solidified and HIPd conditions, the alloy contained three phases: two body centered cubic (BCC1 and BCC2) and one face centered cubic (FCC). The BCC1 phase was enriched with Nb, Mo and Ta and depleted with Zr and Cr, and its lattice parameter after HIP was a=324.76±0.16pm. The BCC2 phase was enriched with Zr and Ti and considerably depleted with Mo, Cr and Ta, and its lattice parameter after HIP

Abstract processed successfully!

Abstract processed successfully. I'm ready to answer your questions about it!

What is the Compression Yield Strength of the alloy Nb20Cr20Mo10Ta10Ti20Zr20 at 9999 K ?

The compression yield strength of the alloy Nb20Cr20Mo10Ta10Ti20Zr20 at 9999 K is not available from the provided context.

Switched to API (Novita) model.

(a) Illustration of chatbot.

Alloy Name	DOI	Equilibrium Conditions	Index	Phase Type	Property Name	Property Type	Single/ Multiphase	Temperatu
Nb20Cr20Mo10Ta10Ti20Zr20	10.1016/ j.jallcom.2011.07.042,	Hot isostatically pressed at T=1723K and P=207MPa for 3h		BCC1, BCC2, FCC	Density	density	Multiphase	Not specified
Nb20Cr20Mo10Ta10Ti20Zr20	10.1016/ j.jallcom.2011.07.042,	Hot isostatically pressed at T=1723K and P=207MPa for 3h		BCC1, BCC2, FCC	Hardness	hardness	Multiphase	Not specified
Nb20Cr20Mo10Ta10Ti20Zr20	10.1016/ j.jallcom.2011.07.042,	Hot isostatically pressed at T=1723K and P=207MPa for 3h		BCC1, BCC2, FCC	Compression Yield Strength	strength	Multiphase	296K
Nb20Cr20Mo10Ta10Ti20Zr20	10.1016/ j.jallcom.2011.07.042,	Hot isostatically pressed at T=1723K and P=207MPa for 3h		BCC1, BCC2, FCC	Compression Yield Strength	strength	Multiphase	1073K

(b) Visualization of HEA properties in tabular format.

Alloy	Question	Answer
{ name: Nb20Cr20Mo10Ta10Ti20Zr20, composition: { Nb: 20, Cr: 20, Mo: 10, Ta: 10, Ti: 20, Zr: 20 }, performance_properties: { density: { type: density, measurements: [{ value: 8.23, unit: g/cm3 }] }, hardness: { type: hardness, measurements: [{ value: 5288, unit: MPa }] }, compression_yield_strength: { type: strength, measurements: [{ temperature: 296K, value: 1595, unit: MPa }, { temperature: 1073K, value: 983, unit: MPa }, { temperature: 1273K, value: 546, unit: MPa }, { temperature: 1473K, value: 171, unit: MPa }] } }, material_conditions: { equilibrium_conditions: Hot isostatically pressed at T=1723K and P=207MPa for 3h, single_or_multiphase: Multiphase, phase_type: BCC1, BCC2, FCC }, doi: 10.1016/ j.jallcom.2011.07.042, }	What is the density of Nb20Cr20Mo10Ta10Ti20Zr20?	8.23 g/ cm^3
{ name: Nb20Cr20Mo10Ta10Ti20Zr20, composition: { Nb: 20, Cr: 20, Mo: 10, Ta: 10, Ti: 20, Zr: 20 }, performance_properties: { density: { type: density, measurements: [{ value: 8.23, unit: g/cm3 }] }, hardness: { type: hardness, measurements: [{ value: 5288, unit: MPa }] }, compression_yield_strength: { type: strength, measurements: [{ temperature: 296K, value: 1595, unit: MPa }, { temperature: 1073K, value: 983, unit: MPa }, { temperature: 1273K, value: 546, unit: MPa }, { temperature: 1473K, value: 171, unit: MPa }] } }, material_conditions: { equilibrium_conditions: Hot isostatically pressed at T=1723K and P=207MPa for 3h, single_or_multiphase: Multiphase, phase_type: BCC1, BCC2, FCC }, doi: 10.1016/ j.jallcom.2011.07.042, }	What is the hardness of Nb20Cr20Mo10Ta10Ti20Zr20?	5288 MPa
{ name: Nb20Cr20Mo10Ta10Ti20Zr20, composition: { Nb: 20, Cr: 20, Mo: 10, Ta: 10, Ti: 20, Zr: 20 }, performance_properties: { density: { type: density, measurements: [{ value: 8.23, unit: g/cm3 }] }, hardness: { type: hardness, measurements: [{ value: 5288, unit: MPa }] }, compression_yield_strength: { type: strength, measurements: [{ temperature: 296K, value: 1595, unit: MPa }, { temperature: 1073K, value: 983, unit: MPa }, { temperature: 1273K, value: 546, unit: MPa }, { temperature: 1473K, value: 171, unit: MPa }] } }, material_conditions: { equilibrium_conditions: Hot isostatically pressed at T=1723K and P=207MPa for 3h, single_or_multiphase: Multiphase, phase_type: BCC1, BCC2, FCC }, doi: 10.1016/ j.jallcom.2011.07.042, }	What is the compression yield strength of Nb20Cr20Mo10Ta10Ti20Zr20 at 296K?	1595 MPa
{ name: Nb20Cr20Mo10Ta10Ti20Zr20, composition: { Nb: 20, Cr: 20, Mo: 10, Ta: 10, Ti: 20, Zr: 20		

(c) Illustration of generated question answer pairs.

Figure 3.5: Implementation of the HEA knowledge system showcasing different views.

Chapter 4

Limitations and Technical Challenges

During the implementation of our system, we encountered various limitations and technical challenges across different project components. Understanding these limitations is crucial for both interpreting our results and guiding future improvements.

4.0.1 Domain-Specific Accuracy Limitations

- **Chemical composition parsing:** Our system achieved only 90% accuracy in correctly parsing complex chemical compositions, particularly struggling with non-standard notation and elemental ratios expressed as ranges or with error margins.

- **Property condition associations:** Accurately linking material properties to their specific measurement conditions proved challenging. In approximately 10% of cases, our system incorrectly associated properties with temperatures or other conditions, particularly when multiple conditions were mentioned in proximity.

- **Handling contradictory information:** When processing multiple abstracts that contained conflicting information about the same alloy, our system lacked sophisticated mechanisms for identifying and resolving these contradictions. We implemented basic conflict reporting, but more advanced reconciliation remained beyond our current capabilities.

- **Limited cross-reference capability:** Our current implementation treats each abstract as an independent unit of information, lacking the ability to draw connections or identify relationships between different papers discussing the same or similar alloys. This limits the system's ability to provide comprehensive synthesis of knowledge across the literature.

4.0.2 Dataset and Knowledge Representation Limitations

- **Limited training data size:** Our HEA_QA_DATASET contained approximately 250 datapoints (from a total of 530 available), which is relatively small for fine-tuning language models. This limited size likely restricted our models' ability to generalize across the full range of HEA-related queries.

- **Technical terminology variations:** We observed inconsistent terminology usage across different research papers. For example, some papers referred to ”yield strength” while others used ”yield stress” for effectively the same property, creating challenges for consistent information extraction.

- **Numerical representation challenges:** Many HEA properties involve complex numerical data with error ranges, temperature dependencies, and conditional values. We found that language models, even after fine-tuning, struggled with precisely representing these numerical relationships, occasionally conflating values from different experimental conditions.

- **Context window constraints:** The scientific literature frequently presents information that exceeds our context window limitations (512 tokens). This forced trade-offs between including complete experimental details versus broader coverage of multiple properties, potentially affecting answer accuracy for complex questions.

4.0.3 Model Fine-tuning Challenges

- **Computational resource limitations:** Fine-tuning Phi-2 (2.7B parameters) required significant computational resources that exceeded our available hardware. While QLoRA partially addressed this, we still encountered periodic out-of-memory errors during training that necessitated further batch size reductions and gradient accumulation steps.

- **Catastrophic forgetting:** During early experiments with T5, we observed that aggressive fine-tuning occasionally led to ”catastrophic forgetting” where the model would excel at domain-specific questions but regress significantly on general language understanding.

- **Hyperparameter optimization constraints:** Due to computational limitations, we could not perform exhaustive hyperparameter tuning. For example, with Phi-2, we could only test a single learning rate (2e-4) and rank value (r=8) for the LoRA adaptation, potentially missing more optimal configurations.

- **Overfitting on limited examples:** With RoBERTa-SQuAD2, we observed early signs of overfitting after just 2-4 epochs, with evaluation metrics improving while training metrics plateaued. This limited our ability to extract maximum benefit from continued training, even though the model appeared not fully adapted to the domain.

4.0.4 Inference Performance Limitations

- **Inference latency:** On CPU infrastructure, our fine-tuned Phi-2 model required approximately 1-1.5 minutes to generate an answer, which exceeds user expectation for interactive applications.While our implementation included an API fallback option, this created inconsistency in answer style and quality.

- **Memory consumption during batch operations:** When processing multiple abstracts simultaneously, we encountered memory spikes up to 7GB, particularly during the embedding generation phase. This limited the scalability of our application on standard deployment environments with memory constraints.

- **Retrieval quality inconsistencies:** Our FAISS-based retrieval system occasionally failed to identify the most relevant contexts, especially when questions contained domain-specific terminology not present in the exact form within the indexed text. Despite implementing adaptive thresholds, approximately 12% of test questions received suboptimal context selection.

- **Answer hallucination under uncertainty:** Even with fine-tuning, our models occasionally produced confident-sounding but factually incorrect responses when the relevant information was absent or ambiguous in the provided context. Our attempts to mitigate this through prompt engineering were only partially successful.

4.0.5 Application Development and Deployment Challenges

- **Flask concurrency limitations:** The default Flask development server does not support true concurrency, which created bottlenecks when multiple users attempted to use the system simultaneously. While we implemented some asynchronous processing for long-running tasks, the fundamental single-threaded nature of the server remained a limitation.
- **Session data persistence issues:** Our implementation of session management using filesystem storage occasionally encountered file locking problems, particularly when sessions were terminated unexpectedly. This sometimes resulted in orphaned data that accumulated until manual cleanup was performed.
- **Model loading time:** The initial loading of our fine-tuned models took approximately 45-60 seconds, creating a significant delay during application startup.
- **Client-side performance variations:** JavaScript-based rendering of large result tables (particularly for multi-abstract processing) performed poorly on less powerful devices. We observed rendering times exceeding 5 seconds for datasets with more than 50 extracted properties on standard laptop configurations.

These limitations highlight opportunities for future work in improving both the technical implementation and the fundamental approaches used in our system. Many of these challenges represent active areas of research in applying language models to specialized scientific domains.

Chapter 5

Conclusion

This project describes a systematic method for extracting, structuring, and analyzing literature related to high-entropy alloy (HEA) using Artificial Intelligence methods. By testing different large language models (LLaMA 3.1 8B and Qwen2.5-1.5B-Instruct) and approaches (retrieval-augmented generation (RAG) and few-shot prompting), we have created an effective system for knowledge extraction and question-answer generation in the field of materials science. The major findings of this work include:

- **Thorough Knowledge Extraction:**
  - We collected and processed scientific documents related to high-entropy alloys (HEAs), transforming abstracts from the Scopus database into machine-readable, annotated data using Doccano with expert input. This enabled structured data extraction in JSON format.
  - Three main approaches were used for knowledge extraction: (1) fine-tuning LLaMA 3.1 8B on annotated abstracts, (2) a RAG-based pipeline using FAISS for contextual retrieval and structured output generation, and (3) few-shot prompting with GPT models for extracting key alloy properties and synthesis details.
  - Model outputs were evaluated using a confusion matrix-based framework against expert-annotated gold standards, and performance metrics (Precision, Recall, F1 Score) were calculated for each model using post-processing to extract relevant JSON segments for accurate comparison.

- **Suggestive Q/A Generation:**
  - Context-rich JSON data was used to prompt Qwen for domain-specific Q-A generation across categories like temperature, units, and material state.
  - A two-stage filtering (heuristic + LLM-based) enhanced accuracy and boosted expert-rated selection performance by 33%.
- **Model Finetuning for HEA Qeustion Answering**
  - We finetuned three language models—T5-Base, RoBERTa-Base-SQuAD2, and Phi-2 (with QLoRA)—on a domain-specific QA dataset (**HEA-QA-DATASET.json**) structured in SQuAD format to enhance performance on HEA literature comprehension tasks.
  - Each model was optimized using distinct strategies: T5 for generative QA, RoBERTa for extractive QA, and Phi-2 with parameter-efficient QLoRA for scalable adaptation to specialized scientific vocabulary and context.
  - The diverse architectures and training techniques offer complementary strengths, providing a robust foundation for accurate information extraction and validation within the HEA research domain.
- **Interactive UI for Scientific Exploration:**
  - A web-based UI allows users to upload or paste abstracts, and view alloy data in tabular and JSON formats.
  - The platform enables Q-A generation and chatbot interaction for on-demand scientific query resolution.

Future Scope: This framework can be expanded to include other domains in materials science, supporting broader knowledge extraction and scientific discovery. With further optimization, the system can improve in speed, scalability, and adaptability—potentially contributing to HEA design, automated literature reviews, and real-time materials recommendation engines.

Bibliography

1. X. Yang, Y. Zhang, and P. K. Liaw, “Microstructure and compressive properties of NbTiVTaAl$_x$ high entropy alloys,” in Materials Science Forum, C. M. Wang and C. J. Peng, Eds., 2012, pp. 292–298.

2. S. Y. Chen, X. Yang, K. A. Dahmen, P. K. Liaw, and Y. Zhang, “Microstructures and crackling noise of Al$_x$NbTiMoV high entropy alloys,” Entropy, vol. 16, pp. 870–884, 2014.

3. C.-M. Lin, C.-C. Juan, C.-H. Chang, C.-W. Tsai, and J.-W. Yeh, “Effect of Al addition on mechanical properties and microstructure of refractory Al$_x$HfNbTaTiZr alloys,” Journal of Alloys and Compounds, vol. 624, pp. 100–107, 2015

4. S. Gorsse, D. B. Miracle, and O. N. Senkov, “Mapping the world of complex concentrated alloys,” Acta Materialia, vol. 135, pp. 177–187, 2017. doi: 10.1016/j.actamat.2017.06.027.

5. D. Qiao, H. Jiang, X. Chang, Y. Lu, and T. Li, “Microstructure and mechanical properties of VTaTiMoAl$_x$ refractory high entropy alloys,” Materials Science Forum, pp. 638–642, 2017.

6. Jha, D., Ward, L., Paul, A., Liao, W.-k., Choudhary, A., Wolverton, C., Agrawal, A. (2018). ElemNet: Deep learning the chemistry of materials from only elemental composition. Scientific Reports, 8(1), 17593. Nature Publishing Group.

7. O. N. Senkov, J. K. Jensen, A. L. Pilchak, D. B. Miracle, and H. L. Fraser, “Compositional variation effects on the microstructure and properties of a refractory

high-entropy superalloy $AlMo_{0.5}NbTa_{0.5}TiZr$," Materials Design, vol. 139, pp. 498–511, 2018

8. X. Li, Y. Sun, and G. Cheng, "TSQA: Tabular scenario-based question answering," in arXiv Computation and Language, Jan. 14, 2021.

9. Schmid, U. (2022). MatSciBERT: A materials domain language model for text mining and information extraction. npj Computational Materials, 8(1).

10. P. Shetty, A. C. Rajan, C. Kuenneth, S. Gupta, L. P. Panchumarti, L. Holm, C. Zhang, and R. Ramprasad, "A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing," in npj Computational Materials, Sept. 29, 2022.

11. Kritesh Kumar Gupta, Surajit Das Barman, S Dey, Susmita Naskar,T. Mukhopadhyay. (2024). On exploiting nonparametric kernel-based probabilistic machine learning over the large compositional space of high entropy alloys for optimal nanoscale ballistics. *Dental Science Reports*, 14(1). Nature Portfolio.

12. A. Chandrasekhar, J. Chan, F. Ogoke, O. Ajenifujah, and A. B. Farimani, "AMGPT: A large language model for contextual querying in additive manufacturing," arXiv preprint arXiv:2405.12345, May 24, 2024.

13. Y. Geng, J. Choi, H. Song, O. Miano, J. S. Choi, K. Bang, B. Lee, S. S. Sohn, D. Buttler, A. M. Hiszpanski, S. S. Han, and D.-H. Kim, "MaTableGPT: GPT-based table data extractor from materials science literature," Cornell University, June 8, 2024.

14. S. Ghosh, N. R. Brodnik, C. Frey, C. S. Holgate, T. M. Pollock, S. Daly, and S. Carton, "Toward reliable ad-hoc scientific information extraction: A case study on two materials datasets," Cornell University, June 8, 2024.

List of Publications based on this research work

1. A. Sirohi, R. SriViswa, S. Gorain, V. Bansal and Dr K. Gupta, "Automated annotation of materials science literature on high-entropy alloys for efficient data extraction," Manuscript in preparation, *yet to be submitted.*