LLM Hackathon for Applications and Materials in Chemistry 2025

HEAQuery - Project Summary

Team Members:

- Taradutt Pattnaik[1]
- Alexander J Horvath[1]
- Sanjeev Nayak[1]

[1]University of Connecticut Dept. of Materials Science & Engineering

Date: 9/12/2025

HEAQuery: LLM-Powered Research Assistant for High-Entropy Alloys

Project Summary

High-entropy alloys (HEAs) are an emerging class of materials with complex compositions and tunable properties, making them a rich subject for both experimental and computational research. Navigating the vast literature and heterogeneous datasets to extract actionable insights, however, is challenging.

HEAQuery is a hybrid intelligent query system developed to address this challenge. By integrating structured HEA datasets, unstructured scientific literature, and large language models (LLMs), HEAQuery enables researchers to ask natural-language questions and retrieve **data-driven answers grounded in both experimental results and published research**. The system combines PDF preprocessing, summarization, vector embeddings, dataset cleaning, semantic search, and LLM reasoning to deliver a user-friendly, interactive interface for exploring high-entropy alloys.

Technical Overview

Step 1: PDF Preprocessing

The first stage of HEAQuery involves preparing a **structured text corpus** from raw PDF open-access research papers on HEAs. A preprocessing script iterates over all PDFs in a designated directory and extracts their text using **PyMuPDF (fitz)**.

Raw PDF text often contains artifacts such as excessive whitespace, URLs, DOIs, or page-number footers. The script applies multiple cleaning steps to standardize the content, making it suitable for downstream natural language processing and embedding generation.

Additionally, a lightweight **metadata extractor** infers the paper title and first author from the initial lines of each document. Cleaned text and metadata are stored in a dictionary indexed by filename and serialized as a pickle file (`raw_corpus.pkl`). This structured corpus forms the foundation for subsequent **chunking, embedding, and retrieval-based querying**.

Step 2: Summarization, Chunking, and Vector Database Construction

The second stage transforms the preprocessed text into a **high-quality, searchable knowledge base**.

1. **Section-level extraction** identifies key scientific sections such as **abstract, introduction, results/discussion, and conclusion**, while discarding acknowledgments, references, and author lists.

2. Each section is summarized using a **GPU-accelerated BART model[4]**, producing concise scientific summaries.

3. Summaries are split into overlapping text chunks using a **recursive text splitter**, preserving context while ensuring manageable chunk size. Each chunk also receives a **one-sentence mini-summary**.

4. Chunks are converted into **LangChain Document objects** and embedded using **MatSciBERT**, a model specialized for materials science. To handle large datasets efficiently, embeddings are processed in batches.

5. Embeddings are stored in a **FAISS vector index**, supporting **Retrieval-Augmented Generation (RAG)**. The index is saved incrementally for reliability.

This stage enables **fast semantic retrieval** from over 3,500 open-access scientific papers, providing the backbone for literature-grounded LLM reasoning.

Step 3: HEA Dataset Cleaning and Standardization

The third stage creates a **unified numerical foundation** from three publicly available HEA datasets[1-3], which originally contained inconsistent column names, variable alloy formula formats, and redundant element-fraction columns.

A consistent cleaning procedure is applied:

- Columns are renamed to **standardized labels**.

- Extraneous formatting is removed.

- Alloy compositions are normalized using a **custom chemical formula parser**, generating a `composition_norm` key for cross-dataset merging.

Only relevant **physical, thermodynamic, microstructural, and metadata fields** are retained:

- **Dataset 1 (MPEA):** Experimental mechanical properties and microstructure
- **Dataset 2:** ML-derived features and design parameters
- **Dataset 3 (ACHIEF):** Thermodynamic and electronic descriptors

The cleaned datasets are saved as `dataset1_clean.csv`, `dataset2_clean.csv`, and `dataset3_clean.csv`. These structured datasets allow HEAQuery to answer **data-driven queries** about HEA compositions, properties, and phase behavior.

Step 4: Integrated HEAQuery System

The final stage integrates all components into a **single interactive application**.

1. The system loads the **cleaned HEA datasets** and the **FAISS vector index** of MatSciBERT[5] embeddings, enabling semantic search across >3,000 research papers.
2. A **query parser** interprets user questions, extracting structural constraints (e.g., FCC/BCC), property thresholds (e.g., hardness > 200), and comparative terms (e.g., highest hardness).
3. Datasets are filtered using **synonym-aware column matching** and numeric comparisons.
4. FAISS retrieves the most relevant literature passages, providing **textual context** for LLM reasoning.
5. Structured data and textual context are input to an **LLM pipeline** (currently GPT-2[6]), which generates **summaries of relevant alloy compositions and properties**.

A **Gradio interface** presents three outputs to the user:

1. **LLM-generated alloy summary**
2. **Merged table of matching alloys** from all datasets
3. **FAISS-retrieved scientific context**

For example, a query such as *“Which HEAs have FCC structure and hardness > 200?”* returns a literature-grounded, data-driven answer, demonstrating HEAQuery’s **hybrid approach combining structured data, unstructured text, and LLM reasoning**.

Conclusion

HEAQuery illustrates how **hybrid AI techniques** can transform the way materials researchers access and analyze high-entropy alloy knowledge. By combining **PDF preprocessing, summarization, semantic embeddings, dataset harmonization, and LLM reasoning**, the system provides a scalable, intelligent query platform. This approach can be extended to other materials science domains, offering a roadmap for integrating structured and unstructured data with AI-driven reasoning.

References

1. MPEA dataset: C. Borg, “Expanded dataset of mechanical properties and observed phases of multi-principal element alloys”. figshare, 12-Jul-2020, doi: 10.6084/m9.figshare.12642953.v9.

2. ML Pred dataset: R. Machaka, G. T. Motsi, L. M. Raganya, P. M. Radingoana, and S. Chikosha, “Machine learning-based prediction of phases in high-entropy alloys: A data article,” Data in Brief, vol. 38, p. 107346, Oct. 2021, doi: https://doi.org/10.1016/j.dib.2021.107346.

3. Achief dataset: C. E. Precker, A. Gregores Cotoand S. Muíños Landín, “Materials for Design Open Repository. High Entropy Alloys”. Zenodo, Aug. 03, 2021. doi: 10.5281/zenodo.5155150.

4. MatSciBert: Gupta, T., Zaki, M., Krishnan, N. M. A., & Mausam. (2022). MatSciBERT: A materials domain language model for text mining and information extraction. *Npj Computational Materials*, *8*(1). https://doi.org/10.1038/s41524-022-00784-w

5. BART-CNN: Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., … Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1910.13461

6.GPT-2: Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.