LLM Hackathon for Applications and Materials in Chemistry 2025

HEA Query - Project Summary

Project Name: HEA Query

Team Members:

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

1University of Connecticut Dept. of Materials Science & Engineering

Date: 9/12/2025

Summary:

HEA Query is an LLM-powered research assistant for High Entropy Alloys (HEAs), enabling intelligent access to both unstructured literature and structured experimental datasets.

We began by curating a large open-access corpus of ~3500 HEA-related papers. These were parsed into logical sections (abstract, methods, etc.), semantically chunked, and indexed using FAISS with BAAI/bge-base-en embeddings.

In parallel, we cleaned and harmonized three well-known HEA datasets containing alloy compositions and their associated physical or thermodynamic properties (e.g., hardness, strength, phase, mixing enthalpy). Each dataset was normalized and mapped to a canonical schema.

To support natural querying, we integrated:

- Semantic search over literature (FAISS)
- Rule-based filtering of structured datasets
- LLM-powered response generation using Mistral-7B

The result is a unified system that can answer domain-specific questions like:

“List alloys with FCC phase and HV > 200”

Our interactive **Gradio app** combines natural language understanding with tabular results and scientific paper snippets, making HEA research both faster and more insightful.

Technical Overview:

Resource 1: Literature Corpus Processing

- **Data Source**: Open-access PDFs (~3,500 papers) related to High Entropy Alloys (HEAs).

- **Text Extraction**:
  - Used PyMuPDF to extract raw text from PDF pages.
  - Performed **deduplication** using MD5 hashing to skip repeated documents.
- **Section Parsing**:
  - Extracted structured sections from raw text using regex:
    - abstract, introduction, methods, conclusion
- **Chunking**:
  - Applied RecursiveCharacterTextSplitter from LangChain to split sections into manageable semantic chunks.
  - Chunk size: 500 tokens with 50-token overlap.
- **Embedding + Indexing**:
  - Embedded using **BAAI/bge-base-en** model via HuggingFaceEmbeddings.
  - Indexed using **FAISS** (batch-wise, with intermediate saving).
  - Result: Searchable vector database of paper chunks.

Resource 2: Structured HEA Datasets

- Cleaned and normalized **three CSV datasets** on HEAs:
  - **MPEA Dataset**: Experimental data (density, modulus, grain size, etc.)
  - **ML Pred Dataset**: Design parameters + predicted properties (Hmix, Smix, etc.)
  - **Achief Dataset**: Thermodynamic and structural descriptors (Tm, VEC, phase, etc.)
- Applied:
  - **Column renaming** for consistency.
  - **Formula normalization** via regex (e.g., sorting elements, removing spaces).
  - **Dropped irrelevant element-fraction columns**.
- All cleaned datasets saved in /hea_datasets

LLM Setup

- Loaded **Mistral-7B-Instruct v0.3** (via Hugging Face) with:
  - Automatic device mapping (torch_dtype=torch.float16)
  - Run via transformers.pipeline("text-generation")

CSV + FAISS Query Intelligence

- **Synonym Mapping**:
  - Handled multiple naming conventions (e.g., "HV", "Vickers hardness" → hardness)
- **CSV Filtering**:

- Parsed numeric queries like HV > 200, YS < 1000 MPa.
- Filtered categorical attributes like phase structure: FCC, BCC, etc.
- Matched entries from each dataset and returned up to 10 rows per dataset.

- **FAISS Search**:
  - Queried the embedded document corpus using semantic similarity (top_k = 5).
- **Prompt Construction**:
  - Combined relevant paper text (FAISS) + matching dataset rows into a **unified prompt**.
  - Used the Mistral model to generate natural language answers.

Interactive Gradio App

- Built a 3-pane app using **Gradio**:
  - **LLM Answer**: Natural language explanation/summary.
  - **CSV Matches**: Tabular preview of matched alloys from datasets.
  - **FAISS Paper Context**: Raw chunk text from relevant papers
- App title: **"HEA Query"**
- Description: Supports domain-specific queries across >250,000 paper chunks and 3 structured datasets.

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. *Mistral 7B* (2023). "Mistral 7B" — Mistral AI. arXiv:2310.06825.

5. *Mixtral of Experts (MoE)* (2023). "Mixtral of Experts" — Mistral AI. arXiv:2312.17263