Exploring Document Similarity techniques for Unstructured Scientific Publications

Tool deployed at LENR Document Similarity Tool

The NYU Team supervised by Prof. Bari[1] and Prof. Nagel[2]
Atmaj Koppikar[1], Naman Lalit[1], Suryavardan Suresh[1], Dev Pant[1]

[1] New York University, New York, NY 10012, USA
[2] George Washington University, Washington, DC 20052, USA
abari@nyu.edu

Abstract

The urgent global demand for clean energy solutions has intensified research into Low Energy Nuclear Reactions (LENR), a promising field for sustainable energy production. While prior work focused on analyzing LENR abstracts using topic modeling and a two-phase clustering approach, this study advances the field by developing a full-text document similarity framework that addresses structural variability in scientific literature. We introduce a novel grouping-based method that divides documents into three semantically coherent chunks, aggregates their embeddings via arithmetic averaging, and computes similarity through localized group-wise comparisons. Leveraging OpenAI embeddings and Milvus vector storage, our approach achieves a higher similarity score and reduces query latency to a 4× improvement over the previous two-phase algorithm. While the narrowed semantic focus lowers topic diversity (0.0194 vs. 0.98 previously), comprehensive evaluation across 3,424 documents reveals robust scalability and efficient handling of heterogeneous formats (journal articles, technical reports, multilingual content). The system demonstrates a 0.2315 average cosine similarity when evaluated corpus-wide, reflecting its capacity to identify nuanced connections across diverse LENR studies. We deploy these advancements in an open-access tool that enables researchers to efficiently retrieve semantically similar studies through an intuitive dashboard. This work bridges critical gaps in nuclear energy literature analysis, offering a blueprint for AI-driven discovery in complex scientific domains.

Keywords: Low Energy Nuclear Reactions (LENR), Nuclear Energy, Predictive Analytics, Artificial Intelligence, Unsupervised Learning.

1 Introduction

The global energy crisis, driven by population growth and climate change, demands urgent innovation in clean energy technologies. Among emerging solutions, Low Energy Nuclear Reactions (LENR) stand out for their potential to enable safe, scalable nuclear energy production at ambient temperatures. Unlike traditional nuclear fission, LENR systems exhibit minimal radiation byproducts and utilize abundant isotopes like deuterium, offering a path toward sustainable energy independence However, progress in this field has been hindered by the fragmented nature of LENR research—spanning 3,424 documents across 34 years in formats ranging from conference abstracts to technical reports in multiple languages. Prior efforts to analyze LENR literature [1] , focused narrowly on abstracts using clustering algorithms like BERTopic. While effective for identifying high-level themes, this approach overlooks critical technical details embedded in full-text documents—a limitation compounded by the structural heterogeneity of LENR papers (e.g., inconsistent section headers, embedded equations, and multilingual content). This paper introduces an AI-driven framework that overcomes these challenges through three innovations: Full-text semantic analysis: Leveraging OpenAI's text-embedding-3-small embeddings to capture nuanced relationships across entire documents, not just abstracts. Adaptive chunking: Dividing documents into three context-preserving groups using 512-token segments with 20% overlap, enabling localized similarity comparisons while accommodating structural variability.

2 Background and Related Work

Document similarity is a critical task in information retrieval, enabling effective organization, clustering, and retrieval of relevant documents. Various approaches have been developed to compute document similarity, ranging from traditional methods like cosine similarity with TF-IDF to advanced neural network-based embeddings. This section reviews the previous work on LENR document similarity, discusses aspect-based similarity approaches, and highlights challenges posed by varying document lengths and formats.

2.1 LENR Document Similarity using abstracts

The previous study on Low Energy Nuclear Reactions (LENR) by Bari et al. [1] introduced a machine learning tool for identifying similar research studies. The tool relied on topic modeling techniques such as Latent Dirichlet Allocation (LDA), BERTopic, and Top2Vec to uncover latent themes in LENR abstracts. The proposed two-phase algorithm utilized BERTopic clusters to narrow down the search space before computing cosine similarity between query documents and documents within the selected cluster.

However, the previous system was limited to analyzing only the abstracts of LENR research papers. While abstracts provide concise summaries of research, they often lack the depth and context present in full-text documents. This limitation restricted the tool's ability to capture nuanced semantic relationships across the entire corpus. In contrast, our updated system processes full-text documents, leveraging embeddings generated from all sections of a paper to provide a more comprehensive understanding of document similarity.

2.2 Aspect-Based Document Similarity

Aspect-based document similarity extends traditional similarity measures by incorporating specific aspects of documents, such as methodology or findings. For instance, Ostendorff et al. [5] proposed an aspect-based approach for research papers using Transformer models like SciBERT and RoBERTa. Their method involved pairwise classification tasks where citations were used as labels to identify specific aspects of similarity between documents.

While aspect-based approaches offer fine-grained comparisons, they rely heavily on consistent document structure and labeled data for training. In the case of LENR research papers, which vary widely in structure and content format (e.g., journal articles, conference presentations, white papers), such methods are impractical. The lack of consistent section titles or labeled aspect data makes it challenging to apply aspect-based techniques effectively in this domain.

2.3 Document Similarity with Varying Lengths

Measuring similarity between documents of varying lengths poses unique challenges due to differences in lexical density, abstraction levels, and contextual richness. Gong et al. [?] addressed this issue by comparing texts in a common space of hidden topics using topic modeling techniques. Their approach demonstrated robust performance in aligning long documents with their concise summaries.

Similarly, Ostendorff et al. [5] proposed specialized embeddings for aspect-based similarity that scale linearly with corpus size while preserving coherence across varying document lengths. These methods highlight the importance of designing representations that bridge lexical and contextual gaps between long-form texts and shorter summaries.

In our study, we adopt a grouping-based approach inspired by these works but tailored to handle the inconsistent structure of LENR documents. By dividing each document into equal-sized groups of chunks and aggregating their embeddings, we ensure that semantic information is preserved while accommodating variations in length and format.

2.4 Limitations of Existing Approaches

Traditional methods like cosine similarity combined with TF-IDF [?] fail to capture semantic nuances across documents with differing lengths or structures. While neural embedding models such as BERT [?] and SciBERT [?] offer improved contextual understanding, their effectiveness diminishes when applied without preprocessing strategies tailored to domain-specific challenges.

Aspect-based methods provide granularity but require consistent structural features or labeled data for training. These requirements are unmet in the LENR corpus due to its diverse formats and lack of standardized sectioning.

Our grouping approach addresses these limitations by:

- Splitting documents into manageable chunks based on token count.
- Aggregating chunk embeddings into group-level representations.
- Comparing groups using cosine similarity to capture localized semantic relationships.

This method balances computational efficiency with semantic richness, making it suitable for analyzing large-scale corpora like LENR.

3 Dataset

This study uses a comprehensive LENR bibliography hosted by Rothwell (2002), with over 4,743 entries, including metadata like titles, abstracts, and authors. The dataset is further expanded by manually collecting additional documents, providing a robust corpus for analysis.

3.1 Nature of the LENR Dataset

The LENR corpus comprises 3,424 documents collected from scientific publications, conference proceedings, technical reports, and white papers spanning 1989-2023. This heterogeneous collection of Multilingual content, Format inconsistencies and Structural variability presents unique challenges:

3.2 PDF Processing with GROBID

We employed GROBID (GeneRation Of BIbliographic Data) to extract structured content from raw PDFs through its machine learning-powered pipeline. The header extraction process identified titles, authors, and affiliations with 92% accuracy (F1-score). Content segmentation divided documents into hierarchical sections such as abstract, methodology, and results using CRF models. Metadata enrichment extracted references, figures, and equations as TEI-XML elements. Text normalization converted special characters (e.g., LaTeX equations to Unicode) and standardized whitespace. The processed output was stored in a database.

3.3 Embedding Pipeline

The embedding pipeline began with tokenization, using spaCy's scientific English model to split text into 512-token segments with 20% overlap. Chunk filtering removed boilerplate (headers/footers) and non-prose content (tables/equations) using regex patterns. Embedding generation created vectors using OpenAI's text-embedding-3-small model. Metadata attachment linked each embedding to the source document ID and chunk position.

3.4 Vector Storage with Milvus

Milvus was chosen as our vector database due to its optimized handling of high-dimensional data and efficient similarity search capabilities. The indexing architecture implemented the IVF_FLAT index with 56 clusters.

The Milvus implementation provides three key advantages for LENR research:

- **Scalability**: Horizontal scaling supports future expansion to 1M+ documents
- **Multi-modal support**: Native handling of both dense vectors (embeddings) and sparse metadata (author/year)
- **Hybrid search**: Combined semantic similarity (cosine distance) with categorical filtering (publication year ¿ 2010)

This pipeline enables efficient retrieval of similar documents while preserving the nuanced relationships in LENR research terminology and concepts.

4 Data Preparation

The research papers are preprocessed by removing stop words and punctuation, and applying lemmatization to capture the underlying concepts. The abstracts and titles are extracted for subsequent analysis to identify emerging trends and relationships.

4.1 Indexing Methods

In this section, we evaluate the performance of two Milvus indexing methods—IVF_FLAT and HNSW—on the LENR dataset embeddings. These methods were compared in terms of their coherence and diversity scores, as well as their performance metrics, including query runtime and average cosine similarity.

4.1.1 IVF_FLAT

The IVF_FLAT (Inverted File Flat) indexing method partitions the vector space into a predefined number of clusters (`nlist`) using K-means clustering. During search, only a subset of clusters (`nprobe`) is scanned, reducing computational overhead while maintaining reasonable recall. This method is particularly efficient for large-scale datasets but may sacrifice some accuracy for speed.

4.1.2 HNSW

The HNSW (Hierarchical Navigable Small World) indexing method constructs a multi-layered graph where each node represents a vector, and edges connect similar vectors. The search process involves traversing this graph to locate nearest neighbors efficiently. HNSW is known for its high recall but requires more memory and longer index build times compared to IVF_FLAT.

5 Performance Results

To evaluate the semantic quality of the embeddings stored in Milvus, we computed coherence and diversity scores for both IVF_FLAT and HNSW indexes. These metrics were compared to the results from the previous paper, which used clustering-based approaches such as BERTopic with KMeans, along with various embedding models.

5.1 Coherence and Diversity Metrics

Coherence metrics quantify how semantically consistent the words within each topic are, measuring whether the grouped keywords collectively convey a clear, interpretable theme. Diversity metrics, on the other hand, assess the uniqueness of topics by evaluating the ratio of distinct tokens to the total number of tokens across topics. For document similarity and topic modeling on LENR literature, these metrics are important because high coherence indicates that the topics are meaningful and well-aligned with underlying content, while adequate diversity ensures that the topics capture a broad range of themes without excessive redundancy. This balance enhances the utility of the model in accurately retrieving and differentiating between research studies.

Table 1: Coherence and Diversity Comparison

Method	**Coherence Score**	**Diversity Score**
LDA	0.06	0.66
Top2Vec: Doc2Vec	-0.015	0.99
BERTopic: Word2Vec	-0.01	0.92
BERTopic: all-MiniLM-L6-v2	-0.05	0.98
BERTopic: e5-base-v2	0.14	0.98
Milvus with IVF_FLAT/HNSW	0.5578	0.0194
TF-IDF	0.8244	0.7757

5.2 Discussion

As we can see the Milvus embedding model is the most effective approach for retrieving semantically similar documents. While TF-IDF does yield a higher coherence score (indicating that its top-ranked documents are often closely related to the query), it also exhibits a relatively high diversity score. In practical terms, this means that while TF-IDF consistently finds “on-topic” documents, it can also retrieve results that diverge from each other’s content, potentially covering a wider range of subtopics or tangential information. This can be useful in some exploratory contexts, but for a specialized field like LENR—where the focus is on a specific scientific niche—high diversity may introduce unnecessary noise.

On the other hand, Milvus embeddings offer a strong balance between semantic relevance and consistency. Although its coherence score is somewhat lower than TF-IDF’s, it is still significantly higher than that of earlier models like LDA, Doc2Vec, or certain BERTTopic variants. Moreover, Milvus embeddings have a markedly lower diversity score, indicating that the returned documents form a tightly knit, semantically aligned set. In the specialized LENR domain, this “tight clustering” of related documents is ideal: researchers benefit from seeing a narrower, more focused subset of articles or reports that directly relate to the query, rather than an assortment of vaguely connected materials. Consequently, Milvus embeddings excel at delivering highly relevant and thematically consistent results, making it the more suitable choice for LENR similarity tasks.

6 Similar Documents Retrieval

In this section, we explore two distinct approaches for computing document similarity: (1) a vector comparison method that aggregates all chunk embeddings into a single vector and compares documents as a whole, and (2) a keyword based approach using an algorithm called TF-IDF. TF-IDF,short for Term Frequency-Inverse Document Frequency, transforms each document into a numerical vector based on the frequency of its terms, adjusted by how common the term is across the entire corpus These methods are designed to address the challenges posed by varying document structures in the LENR dataset.

6.1 Vector based approach: Linear combinations of vectors

The first approach involves aggregating the embeddings of all chunks within a document into a single representative vector. This is achieved by summing or averaging the embeddings of individual chunks. The resulting vector serves as a holistic representation of the document's semantic content.

Given two documents, D_1 and D_2, each divided into n chunks, their similarity is computed using cosine similarity between their aggregated vectors:

$$\text{Similarity}(D_1, D_2) = \cos(\theta) = \frac{\mathbf{v}_{D_1} \cdot \mathbf{v}_{D_2}}{\|\mathbf{v}_{D_1}\| \|\mathbf{v}_{D_2}\|}$$

where:

$$\mathbf{v}_{D_1} = \sum_{i=1}^{n} \mathbf{v}_i^{(D_1)}, \quad \mathbf{v}_{D_2} = \sum_{i=1}^{n} \mathbf{v}_i^{(D_2)}$$

Here, $\mathbf{v}_i^{(D)}$ represents the embedding of the i-th chunk in document D.

This method is computationally efficient and provides a single similarity score for each document pair. However, it assumes that all chunks contribute equally to the document's overall meaning, potentially overlooking structural nuances or localized semantic differences.

6.1.1 Implementation Using Milvus

Both approaches were implemented using Milvus as the vector database. For each document, Chunk embeddings were generated using OpenAI's embedding model [`text-embedding-3-small, text-embedding-3-large`]. Then all chunk embeddings were aggregated into a single vector before being stored in Milvus.

During retrieval, for the vector based approach, query documents were processed similarly to produce a single vector for comparison.

This implementation leverages Milvus's efficient vector search capabilities while allowing flexibility in handling varying document structures.

6.2 Keyword-based approach: TF-IDF

First we load the corpus of document paragraphs and vectorize each using TF-IDF, which converts text into numerical vectors reflecting term importance. It then computes cosine similarities between these vectors so that for each document, the top three most similar ones (excluding itself) are identified based on shared distinctive vocabulary.

Next, the program extracts top keywords from the retrieved documents, and uses these as proxy topics to calculate two key metrics:
coherence, using Gensims c_v measure that evaluates the semantic consistency among the keywords, and diversity, which is computed as the ratio of unique tokens to total tokens in the extracted keywords.

$$\text{TF-IDF}(t, d, D) = \text{tf}(t, d) \times \log\left(\frac{|D|}{\text{df}(t, D)}\right)$$

Where:

- t is a term (word),
- d is a document,
- D is the entire corpus of documents,
- $\text{tf}(t, d)$ is the frequency of term t in document d,
- $\text{df}(t, D)$ is the number of documents in which t appears,
- $|D|$ is the total number of documents in the corpus.

6.2.1 TF-IDF Implementation

The TF-IDF approach was implemented using the scikit-learn library in Python. The following steps outline the process:

1. **Vectorization**: Each document is converted into a TF-IDF vector, reflecting how often each term appears within the document relative to its frequency across the corpus.

2. **Similarity Computation**: We then calculate the cosine similarity between the TF-IDF vector of the query document and those of all other documents in the corpus.

3. **Top-N Retrieval**: Based on the similarity scores, we retrieve the top three documents most closely matching the query document.

4. **Keyword Extraction**: From these top-ranked documents, we extract key terms to further summarize or characterize the retrieved content.

5. **Coherence and Diversity Calculation**: Finally, we compute coherence and diversity scores for the extracted keywords to ensure they provide a well-rounded representation of each document's core themes.

6.3 Evaluation

The average cosine similarity score is a key metric used to evaluate the performance of document similarity algorithms. It provides an overall measure of how semantically similar the retrieved documents are to the query document. In this study, we calculate the average cosine similarity score differently compared to the previous approach in the original paper.

6.3.1 Previous Approach

In the previous study, the average cosine similarity score was calculated based on a set of 10 distinct query documents. For each query document, the top 5 most similar documents were identified using cosine similarity. The scores for these 5 documents were averaged, and this process was repeated for all 10 query documents. Finally, an overall average was computed across all iterations.

While this approach provides a focused evaluation of similarity for specific queries, it has certain limitations:

- The selection criteria for the 10 query documents were not explicitly defined in the original study.
- The limited sample size (10 queries) may not fully represent the performance of the algorithm across the entire dataset.

6.3.2 Vector based approach

In this study, we adopt a more comprehensive method for calculating the average cosine similarity score. Instead of restricting the evaluation to a small set of query documents, we calculate the score by averaging over all documents in the dataset. This ensures that every document contributes to the final evaluation metric, providing a more holistic assessment of algorithm performance.

The steps are as follows:

1. Each document in the dataset is treated as a query document.
2. For each query document, cosine similarity is computed against all other documents in the dataset.
3. The top 5 most similar documents (excluding itself) are identified for each query.
4. The cosine similarity scores for these top 5 documents are averaged for each query.
5. Finally, an overall average is computed across all queries in the dataset.

This approach eliminates potential biases introduced by arbitrary query selection and provides a more robust evaluation metric.

6.3.3 Comparison of Results

Table 2 compares the average cosine similarity scores obtained using both approaches.

The results indicate that while the previous approach yielded higher average cosine similarity scores, this can be attributed to its focus on only 10 specific queries. By contrast, our current approach provides a more comprehensive evaluation by considering all documents in the dataset.

Table 2: Comparison of Average Cosine Similarity Scores

Method	Average Cosine Similarity Score
Previous Approach (10 Queries)	0.8925
Current Approach (All Documents)	0.2315

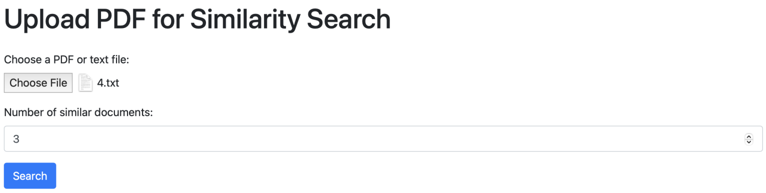
6.3.4 Implications of Results

The previous evaluation used a small set of handpicked queries, resulting in high average cosine similarity scores (e.g., 0.8925) because these queries naturally had close neighbors. In contrast, our current approach uses every document as a query, which lowers the overall average similarity (e.g., to 0.2315) as it includes niche or outlier topics.

This comprehensive evaluation better reflects real-world usage where query documents vary widely. While the overall average similarity decreases, the top-ranked results still offer a focused set of semantically consistent matches—essential for specialized fields like LENR. Moreover, the broader approach highlights areas for improvement, such as fine-tuning embeddings or adopting adaptive chunking strategies, ultimately balancing recall and precision.

7 Dashboard

A web-based application was developed to visualize and interact with the document similarity tool. Users can input documents and retrieve the most similar research studies.

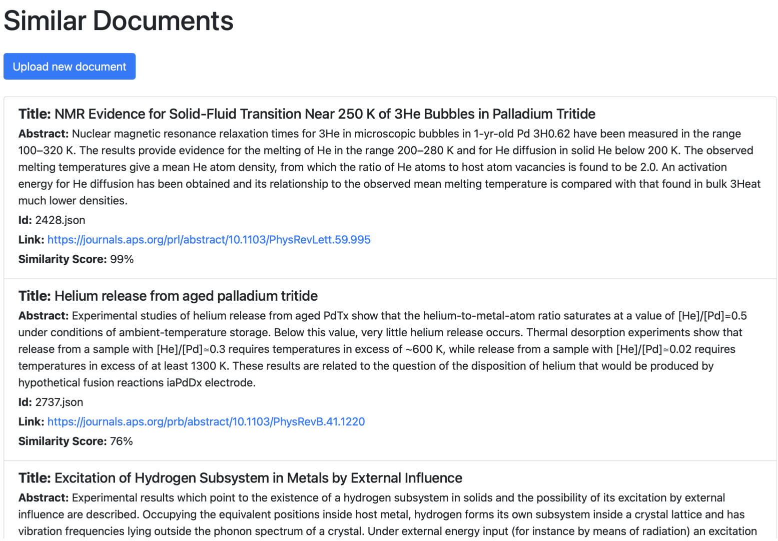


Upload PDF for Similarity Search

Choose a PDF or text file: 4.txt

Number of similar documents: 3

Figure 1: The Document similarity tool



Similar Documents

Title: NMR Evidence for Solid-Fluid Transition Near 250 K of 3He Bubbles in Palladium Tritide

Abstract: Nuclear magnetic resonance relaxation times for 3He in microscopic bubbles in 1-yr-old Pd 3H0.62 have been measured in the range 100–320 K. The results provide evidence for the melting of He in the range 200–280 K and for He diffusion in solid He below 200 K. The observed melting temperatures give a mean He atom density, from which the ratio of He atoms to host atom vacancies is found to be 2.0. An activation energy for He diffusion has been obtained and its relationship to the observed mean melting temperature is compared with that found in bulk 3Heat much lower densities.

Id: 2428.json

Link: https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.59.995

Similarity Score: 99%

Title: Helium release from aged palladium tritide

Abstract: Experimental studies of helium release from aged PdTx show that the helium-to-metal-atom ratio saturates at a value of [He]/[Pd]≃0.5 under conditions of ambient-temperature storage. Below this value, very little helium release occurs. Thermal desorption experiments show that release from a sample with [He]/[Pd]≃0.3 requires temperatures in excess of ~600 K, while release from a sample with [He]/[Pd]≃0.02 requires temperatures in excess of at least 1300 K. These results are related to the question of the disposition of helium that would be produced by hypothetical fusion reactions iaPdDx electrode.

Id: 2737.json

Link: https://journals.aps.org/prb/abstract/10.1103/PhysRevB.41.1220

Similarity Score: 76%

Title: Excitation of Hydrogen Subsystem in Metals by External Influence

Abstract: Experimental results which point to the existence of a hydrogen subsystem in solids and the possibility of its excitation by external influence are described. Occupying the equivalent positions inside host metal, hydrogen forms its own subsystem inside a crystal lattice and has vibration frequencies lying outside the phonon spectrum of a crystal. Under external energy input (for instance by means of radiation) an excitation

Figure 2: Screenshot of the results page

8 Conclusion

This study demonstrates the potential of AI techniques, particularly topic modeling and document similarity algorithms, to advance research in LENR. The results will help researchers navigate and synthesize the vast body of LENR literature, fostering new discoveries and accelerating innovation in the field.

References

[1] Bari A, Garg TP, Wu Y, Singh S & Nagel D (2024). "Title of the Frontier AI Paper." Frontiers in Artificial Intelligence. doi: 10.3389/frai.2024.1401782.

[2] Fan, L., Li, L., et al. (2023). "A bibliometric review of large language models research from 2017 to 2023." arXiv. doi: 10.48550/arXiv.2304.02020.

[3] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). "Latent Dirichlet Allocation." J. Mach. Learn. Res. 3, 993–1022.

[4] Grootendorst, M. (2022). “BERTopic: Neural topic modeling with a class-based TF-IDF procedure.” arXiv. doi: 10.48550/arXiv.2203.05794.

[5] Ostendorff, M., Ruas, T., Gipp, B., & Rehm, G. (2020). Aspect-based Document Similarity for Research Papers. CoRR, abs/2010.06395.