HEAL: Hierarchical Embedding Alignment Loss for Improved Retrieval and Representation Learning

Manish Bhattarai[1], **Ryan Barron**[1,4], **Maksim Eren**[2],
Minh Vu[1], **Vesselin Grantcharov**[1], **Ismael Boureima**[1], **Valentin Stanev**[3],
Cynthia Matuszek[4], **Vladimir Valtchinov**[5], **Kim Rasmussen**[1], **Boian Alexandrov**[1]

[1]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM, USA
[2]Analytics Division, Los Alamos National Laboratory, Los Alamos, NM, USA
[3]Department of Material Science & Engineering, University of Maryland, College Park, MD, USA
[4]Department of Computer Science, University of Maryland, Baltimore County, MD, USA
[5]Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston MA

Correspondence: ceodspspectrum@lanl.gov

Abstract

Retrieval-Augmented Generation (RAG) enhances Large Language Models (LLMs) by integrating external document retrieval to provide domain-specific or up-to-date knowledge. The effectiveness of RAG depends on the relevance of retrieved documents, which is influenced by the semantic alignment of embeddings with the domain's specialized content. Although full fine-tuning can align language models to specific domains, it is computationally intensive and demands substantial data. This paper introduces **H**ierarchical **E**mbedding **A**lignment **L**oss (HEAL), a novel method that leverages hierarchical fuzzy clustering with matrix factorization within contrastive learning to efficiently align LLM embeddings with domain-specific content. HEAL computes level/depth-wise contrastive losses and incorporates hierarchical penalties to align embeddings with the underlying relationships in label hierarchies. This approach enhances retrieval relevance and document classification, effectively reducing hallucinations in LLM outputs. In our experiments, we benchmark and evaluate HEAL across diverse domains, including Healthcare, Material Science, Cyber-security, and Applied Maths.

1 Introduction

Large Language Models (LLMs), such as GPT-4 (OpenAI, 2023), have demonstrated exceptional capabilities in natural language understanding and generation. However, LLMs are prone to *hallucinations*, generating plausible but incorrect or nonsensical content (Ji et al., 2023). Retrieval-Augmented Generation (RAG) frameworks (Lewis et al., 2020) mitigate this issue by integrating external knowledge through document retrieval, enhancing the factual accuracy of LLM outputs. A critical component of RAG systems is the embedding model used for document retrieval. Standard embedding models, however, often fail to capture the hierarchical and semantic relationships within domain-specific corpora, leading to suboptimal retrieval and, consequently, increased hallucinations. This issue is particularly pronounced in domains with increased specificity such as Healthcare, Legal sytem, and Scientific research.

Corpus of documents for a specialized domain inherently exhibit a high degree of semantic coherence, presenting an opportunity to align embedding models for retrieving the most contextually relevant information. Hierarchical Non-negative Matrix Factorization (HNMF) (Eren et al., 2023) is a powerful technique for semantically categorizing documents into clusters that exhibit thematic coherence. By grouping documents into hierarchical clusters of supertopics and subtopics, HNMF provides a rich semantic categorization of the corpus, enabling a deeper understanding of document relationships. Leveraging this semantic knowledge in the form of hierarchical cluster labels, we can align embedding models to preserve hierarchical information within the embedding space. This alignment enhances the embeddings to capture both coarse-grained and fine-grained document similarities, improving contextual relevance in retrieval tasks and enabling better downstream capabilities.

To tackle the challenges of hallucination and suboptimal retrieval in RAG systems, we introduce the **Hierarchical Embedding Alignment Loss (HEAL)**, a refined extension of the Hierarchical Multi-label Contrastive Loss (Zhang et al., 2022). HEAL leverages an improved hierarchical weighting scheme to align embeddings more effectively with the underlying hierarchical structure. By incorporating hierarchical label structures, HEAL fine-tunes embedding models to align with document clusters derived from HNMF. The method computes contrastive losses at each hierarchical level, combining them with depth-specific penalties to emphasize distinctions at higher levels of the hierarchy.

2 Related Work

Contrastive learning has become a cornerstone of representation learning, particularly in computer vision and natural language processing. Methods like SimCLR (Chen et al., 2020) and MoCo (He et al., 2020) have achieved state-of-the-art performance in unsupervised settings by learning representations that are invariant to data augmentations. In supervised contrastive learning, Khosla et al. (2020) extended the contrastive loss to utilize label information, improving performance on classification tasks. Similarly, the SciNCL framework employs neighborhood contrastive learning to capture continuous similarity among scientific documents, leveraging citation graph embeddings to sample both positive and negative examples (Ostendorff et al., 2022). However, these methods generally assume flat label structures and do not exploit hierarchical relationships.

Hierarchical classification has been studied extensively, with approaches such as hierarchical softmax (Goodman, 2001) and hierarchical cross-entropy loss (Deng et al., 2014). These methods aim to leverage hierarchical label structures to improve classification efficiency and accuracy. In the context of representation learning, Deng et al. (2011) introduced hierarchical semantic embedding, aligning image embeddings with WordNet hierarchies. More recent works, such as Bertinetto et al. (2020), have explored hierarchical prototypes to capture hierarchical relationships. Zhang et al. (2022) propose a hierarchical multi-label contrastive learning framework that preserves hierarchical label relationships through hierarchy-preserving losses. Their method excels in scenarios with hierarchical multi-label annotations, such as biological or product classifications. In contrast, our approach focuses on enhancing information retrieval to mitigate hallucinations.

RAG frameworks combine retrieval models with generative models to enhance the factual accuracy of language generation (Lewis et al., 2020). These systems rely heavily on the quality of the embeddings used for retrieval. Prior work has focused on improving retrieval through better indexing and retrieval algorithms (Karpukhin et al., 2020), but less attention has been given to aligning embeddings with hierarchical document structures.

3 Method

In this section, we propose an embedding alignment framework comprising hierarchical label extraction with HNMF, embedding alignment using HEAL, and retrieval with aligned embeddings as outlined in Figure 1.

3.1 Hierarchical Document Clustering with HNMFk.

Hierarchical Non-negative Matrix Factorization with automatic latent feature estimation (HNMFk) Eren et al. (2023) is an advanced technique for uncovering hierarchical patterns within document collections. It builds on traditional Non-negative Matrix Factorization (NMF) Vangara et al. (2021) by dynamically and automatically determining the optimal number of latent features at each level. Effective contrastive learning relies on well-separated document cluster labels to align embeddings effectively. HNMFk's ability to automatically balance stability and accuracy using a bootstrap approach enhances the quality of clustering results. In this work, we utilize the publicly available HNMFk implementation from the TELF library [1].

Given a Term Frequency-Inverse Document Frequency (TF-IDF) matrix $\mathbf{X} \in R^{n \times m}$, where n represents the vocabulary size and m denotes the number of documents, HNMFk performs a sequence of matrix factorizations across hierarchical levels to capture the nested structure of topics. At each level l, the factorization is expressed as $\mathbf{X} \approx \mathbf{W}^{(l)}\mathbf{H}^{(l)}$, where $\mathbf{W}^{(l)} \in R^{n \times k_l}$ is the basis matrix representing latent topics, and $\mathbf{H}^{(l)} \in R^{k_l \times m}$ is the coefficient matrix quantifying the contribution of each topic to the composition of documents. Here, k_l is the number of topics at level l, which is determined automatically through stability analysis (Vangara et al., 2021). This analysis involves bootstrapping the data to create resampled versions of the TF-IDF matrix, applying NMF across a range of k values, and evaluating the stability of clusters across the resampled datasets. The optimal k_l is selected as the value that produces the most consistent clustering results, indicating a robust underlying structure

[1] TELF is available at https://github.com/lanl/T-ELF

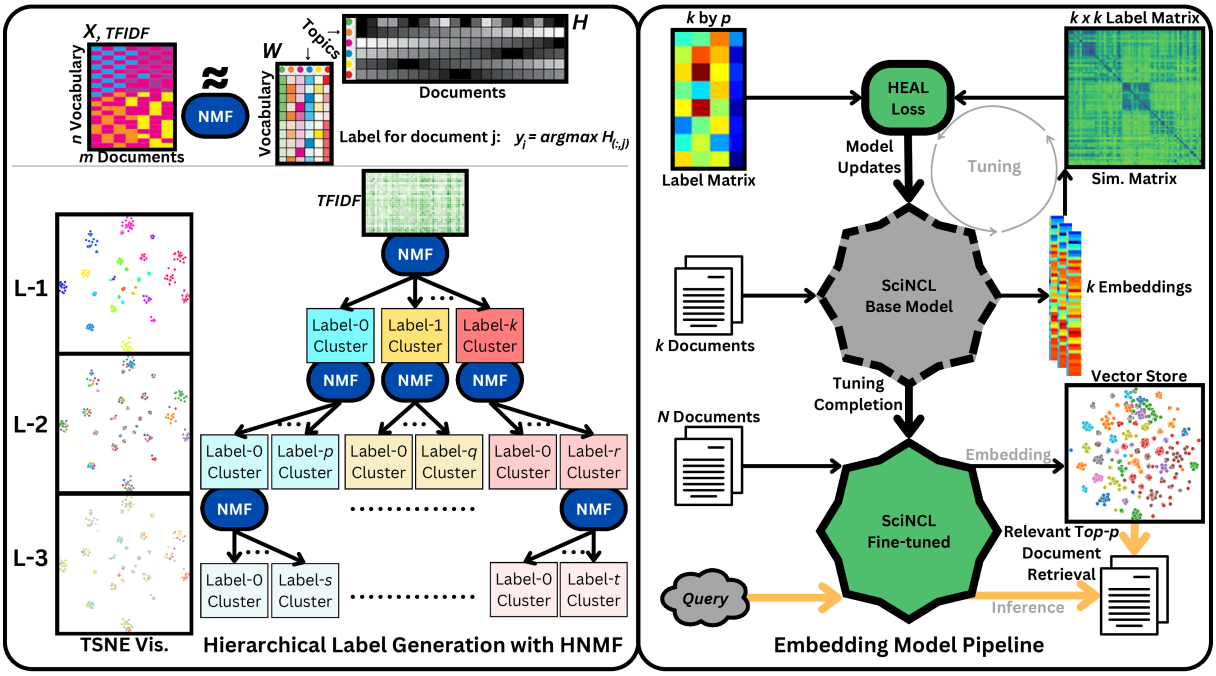


Figure 1: Overview of the HEAL-Based Embedding Model Alignment and Retrieval. *The left side* illustrates hierarchical label generation using HNMF, where documents corresponding to a cluster from each preceding depth are converted into TFIDF matrices and further decomposed to extract sub-clusters. The TSNE visualizations highlighting cluster memberships in document embeddings. *The right side* depicts fine-tuning of the SciNCL model using HEAL loss on generated embeddings and HNMF derived labels. Once trained, the aligned model computes a vector store from the corpus, enabling retrieval of the nearest p documents for a given query embedding.

in the data.

To construct hierarchical labels for each document, the coefficient matrix $\mathbf{H}^{(l)}$ is used to determine topic assignments. For each level l, the topic for document i is identified by selecting the index of the maximum value in the corresponding column of $\mathbf{H}^{(l)}$, expressed as $y_i^{(l)} = \arg\max_k \mathbf{H}_{k,i}^{(l)}$. The hierarchical label for document i is then formed by aggregating the topic assignments across all levels, resulting in $\mathbf{y}_i = (y_i^{(0)}, y_i^{(2)}, \dots, y_i^{(L-1)})$. Here, L is the total number of hierarchical levels, or hierarchical depth that is the number of NMFk operations from the first one to the leaf. y_i^l is the label of sample i at level l, with $l = 0$ corresponding to the *shallowest*(most general or root node) level and $l = L - 1$ to the *deepest* (most fine-grained, or leaf node) level.

3.2 Hierarchical Multilevel Contrastive Loss (HEAL)

Upon the unsupervised data decomposition with HNMFk, the datasets have clusters with hierarchical structures. To incorporate such structures, we propose the HEAL, which extends supervised contrastive loss (Khosla et al., 2020) by introducing level-wise contrastive losses and aggregating them with level-specific penalties.

3.2.1 Level-wise Contrastive Loss

For a batch of N samples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, where $\mathbf{x}_i \in R^d$ is the input and $\mathbf{y}_i \in R^L$ is the hierarchical cluster label, we obtain normalized embeddings $\{\mathbf{h}_i\}_{i=1}^N$ using an encoder network $f_\theta(\cdot)$:

$$\mathbf{h}_i = \frac{f_\theta(\mathbf{x}_i)}{\|f_\theta(\mathbf{x}_i)\|_2}, \quad \mathbf{h}_i \in R^d. \tag{1}$$

For a given level l, the set of positive samples for sample i is:

$$P(i,l) = \{p \mid \mathbf{y}_p^l = \mathbf{y}_i^l, p \neq i\}. \tag{2}$$

The contrastive loss at level l for sample i is:

$$\mathcal{L}_{i,l} = \frac{-1}{|P(i,l)|} \sum_{p \in P(i,l)} \log \frac{\exp\left(\mathbf{h}_i^\top \mathbf{h}_p / \tau\right)}{\sum_{a=1}^N \exp\left(\mathbf{h}_i^\top \mathbf{h}_a / \tau\right)}. \tag{3}$$

If $P(i,l)$ is empty (i.e., no positive samples at level l for i), $\mathcal{L}_{i,l}$ is excluded from the total loss.

3.2.2 Aggregating Level-wise Losses with Penalties

To prioritize discrepancies at shallower levels, we assign penalties λ_l to each level l, where shallower levels have higher penalties. The penalties are defined as:

$$\lambda_l = \frac{2^{L-l-1}}{\sum_{k=0}^{L-1} 2^k} = \frac{2^{L-l-1}}{2^L - 1}. \tag{4}$$

The penalties λ_l satisfy:

1. $\lambda_l > \lambda_{l+1}$ for $l = 0, 1, ..., L-2$, i.e., penalties decrease for deeper levels.
2. $\sum_{l=0}^{L-1} \lambda_l = 1$, i.e., the penalties are normalized.

The total HEAL loss is then:

$$\mathcal{L}_{\text{HEAL}} = \frac{1}{N} \sum_{l=0}^{L-1} \lambda_l \sum_{i=1}^{N} \mathcal{L}_{i,l}. \tag{5}$$

Algorithm 1 Computation of HEAL Loss

```
Require: Mini-batch $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, temperature $\tau$, number of levels $L$
Compute embeddings: $\mathbf{h}_i = f_\theta(\mathbf{x}_i)/\|f_\theta(\mathbf{x}_i)\|_2$
Initialize total loss: $\mathcal{L}_{\text{HEAL}} \leftarrow 0$
for $l = 0$ to $L - 1$ do
    Compute penalty $\lambda_l$ using Eq. (4)
    for $i = 1$ to $N$ do
        Determine positive set $P(i, l)$ using Eq. (2)
        if $|P(i, l)| > 0$ then
            Compute $\mathcal{L}_{i,l}$ using Eq. (3)
            Update total loss: $\mathcal{L}_{\text{HEAL}} \leftarrow \mathcal{L}_{\text{HEAL}} + \lambda_l \mathcal{L}_{i,l}$
        end if
    end for
end for
return $\mathcal{L}_{\text{HEAL}}$
```

Algorithm 1 outlines the computation of $\mathcal{L}_{\text{HEAL}}$ for a mini-batch.

3.3 Fine-tuning Embedding Models with HEAL for RAG

To enhance retrieval performance in RAG systems, we fine-tune the embedding model to align with the hierarchical structure of the document corpus. Given a specialized document corpus, we first apply HNMFk (as described in Section 3.1) to the corresponding TF-IDF matrix $\mathbf{X}$ producing hierarchical cluster labels $\mathbf{y}_i = (y_i^{(0)}, y_i^{(2)}, \ldots, y_i^{(L-1)})$ for each document i. Next, we generate embeddings from each document x_i using a pretrained embedding model $f_\theta(.)$. The embedding model is initialized with pre-trained weights and produces normalized embeddings $\mathbf{h}_i \in R^d$ for document i. To align embeddings with the hierarchical structure, we optimize the HEAL presented in 3.3.

The embedding model is trained by minimizing $\mathcal{L}_{\text{HEAL}}$ using gradient-based optimization:

$$\theta^* = \arg\min_\theta \mathcal{L}_{\text{HEAL}},$$

where θ are the parameters of the embedding model $f_\theta(\cdot)$.

After fine-tuning, the updated embeddings $\mathbf{h}_i = f_{\theta^*}(\mathbf{x}_i)$ are used to replace the initial embeddings in the vector store. During inference, a query $\mathbf{q}$ is embedded using $f_{\theta^*}(\cdot)$ as $\mathbf{h}_q = f_{\theta^*}(\mathbf{q})$, and retrieves top p documents based on cosine similarity:

$$\text{Similarity}(\mathbf{q}, \mathbf{x}_i) = \frac{\mathbf{h}_q^\top \mathbf{h}_i}{\|\mathbf{h}_q\|\|\mathbf{h}_i\|}.$$

To maximize retrieval performance in RAG systems, it is essential to align the query embeddings with the hierarchically aligned document embeddings. Since queries are typically shorter and may not capture the full semantic richness of the documents, we need to semantically align queries and documents in the embedding space. To achieve this, we generate question-answer (Q&A) pairs using a language model (e.g., LLaMA-3.1 70B) for each document and leverage HEAL to jointly align both query and document embeddings during training. For each document $\mathbf{x}_i$, we generate a set of queries $\{\mathbf{q}_{i,k}\}_{k=1}^{K_i}$, where K_i is the number of queries generated for document i. Each query $\mathbf{q}_{i,k}$ is associated with the same hierarchical labels $\mathbf{y}_i$ as its source document $\mathbf{x}_i$, since it is derived from the content of $\mathbf{x}_i$.We extend the HEAL framework to include both documents and queries by defining a unified set of samples:

$$\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \cup \{\mathbf{q}_{i,k} \mid i = 1, \ldots, N;\ k = 1, \ldots, K_i\}.$$

Each sample $\mathbf{s}_j \in \mathcal{S}$ has an associated hierarchical label $\mathbf{y}_j$, where:

$$\mathbf{y}_j = \begin{cases} \mathbf{y}_i, & \text{if } \mathbf{s}_j = \mathbf{x}_i \text{ (document)} \\ \mathbf{y}_i, & \text{if } \mathbf{s}_j = \mathbf{q}_{i,k} \text{ (query generated from document } \mathbf{x}_i\text{).} \end{cases} \tag{6}$$

Based on this dataset, the HEAL is leveraged to finetune the embedding model .

4 Experiments

4.1 Datasets

We evaluate our method on datasets specifically constructed from scientific publications in the domains of Material Science, Medicine, Tensor Decomposition, and Cybersecurity. To construct our datasets, we leveraged the Bibliographic Utility Network Information Expansion (BUNIE) method, a machine learning-based approach that integrates subject-matter expertise in a human-in-the-loop framework (Solovyev et al., 2023). For completeness, we briefly summarize the BUNIE approach in this paper. BUNIE begins with a small core corpus of documents selected by subject-matter experts (SMEs). From this starting point, it constructs a citation network to identify additional relevant documents, leveraging BERT based text embeddings to assess semantic similarity. Through iterative cycles of dataset expansion and pruning—guided by embedding visualization, topic modeling, and expert feedback—the method ensures the corpus is both comprehensive and domain-specific. We apply this procedure to each scientific domain with guidance from SMEs, who provide target keywords/phrases and/or a core set of papers relevant to the sub-topic of interest within the domain. Using this knowledge base, we employ BUNIE to expand the dataset from the initial core papers to a larger collection of domain-specific documents.

1. **Material Science**: A collection of 46,862 scientific articles, which explore 73 Transition Metal Dichalcogenides (TMD) compounds, combining transition-metal and chalcogen atoms (S, Se, or Te). With a layered structure similar to graphite, TMDs excel as solid lubricants and exhibit unique quantum phases like superconductivity and charge density waves. Their atomically thin layers offer tunable properties, with applications in spintronics, optoelectronics, energy harvesting, batteries, and flexible electronics.

2. **Healthcare**: A collection of 9,639 scientific articles, which examine Pulmonary Hypertension (PH) disease - a rare condition causing elevated pulmonary arterial pressure, right heart strain, and reduced oxygen delivery. The WHO classifies PH into five groups based on causes, including pulmonary arterial hypertension (PAH), which has a prevalence of 15-25 cases per million in the U.S. Treatments such as endothelin receptor antagonists and prostacyclin analogs aim to improve symptoms, but prognosis varies, with untreated PAH having a median survival of less than three years.

3. **Applied Mathematics:** A collection of 4,624 scientific articles, which explore tensor network techniques, such as Tensor-Train (TT) decomposition, which recently emerged as a powerful mathematical tool for solving large-scale Partial Differential Equations (PDEs). Tensor network PDE solvers efficiently manage high-dimensional data by mitigating the curse of dimensionality, drastically reducing computational costs and memory usage while maintaining high solution accuracy. These advancements hold significant promise for breakthroughs in scientific computing, including material science, climate modeling, and engineering design optimization.

4. **Cyber-security**: We created a dataset of 8,790 scientific publications focusing on the application of tensor decomposition methods in cybersecurity and ML techniques for malware analysis. This dataset serves as a knowledge base covering topics for cybersecurity such as ML-based anomaly detection, malware classification, novel malware detection, uncertainty quantification, real-world malware analysis challenges, tensor-based anomaly detection, malware characterization, and user behavior analysis.

4.2 Experimental Setup

For training, we used the Adam optimizer with a learning rate of 10^{-5}, a batch size of 128, and early stopping based on validation performance with a patience of 5 epochs. The experiments were conducted on a high-performance computing cluster, with each node equipped with 4 NVIDIA GH200 GPUs. Document metadata, comprising the title and abstract combined, were used as input. Hierarchical labels were generated using HNMF with dataset-specific factorization depths: Material Science (depth 3), Healthcare (depth 4), Applied Mathematics (depth 3), and Cybersecurity (depth 3). HEAL loss was applied with a temperature parameter of 0.07. The embedding base model, SciNCL (Ostendorff et al., 2022), was chosen for its robust contrastive pretraining on scientific documents, serving as a strong baseline for fine-tuning.

The data was split into 60% training, 20% validation, and 20% test sets, with early stopping monitored on the validation set. Evaluation metrics were reported on the test set, while Q&A retrieval analysis used the entire dataset (train + validation + test) for constructing the vector store.

The efficacy of the RAG system was evaluated at two levels. *First*, we characterized the embeddings on document-level tasks, including hierarchical classification, retrieval, and hallucination measurement. For hierarchical classification, we used a hierarchical classifier applying random forests to each node (Miranda et al., 2023). The classifier is trained on embeddings corresponding to train dataset and evaluated against the test set. We perform this for embeddings derived from aligned and unaligned embedding model. Retrieval performance was assessed by measuring whether retrieved documents belonged to the same hierarchical class as the query document. Hallucination likelihood was evaluated based on the retrieval of incorrect documents for a given query. *Second*, we evaluated the performance of the embedding model within a RAG framework. To support retrieval and hallucination analysis, we used the LLaMA-3.1 70B model to generate 10 Q&A pairs per document using abstracts as input, providing a robust test for embedding alignment and retrieval capabilities. Next, we leveraged the questions as queries to the embedding model to retrieve the best metadata and assessed whether the model retrieved the exact document that generated the query during Q&A analysis, as well as the rank of the returned document within the top 10 results. Furthermore, the retrieved documents were augmented with LLaMA-3.1 70B LLM to generate responses, with hallucinations evaluated based on response accuracy and relevance.

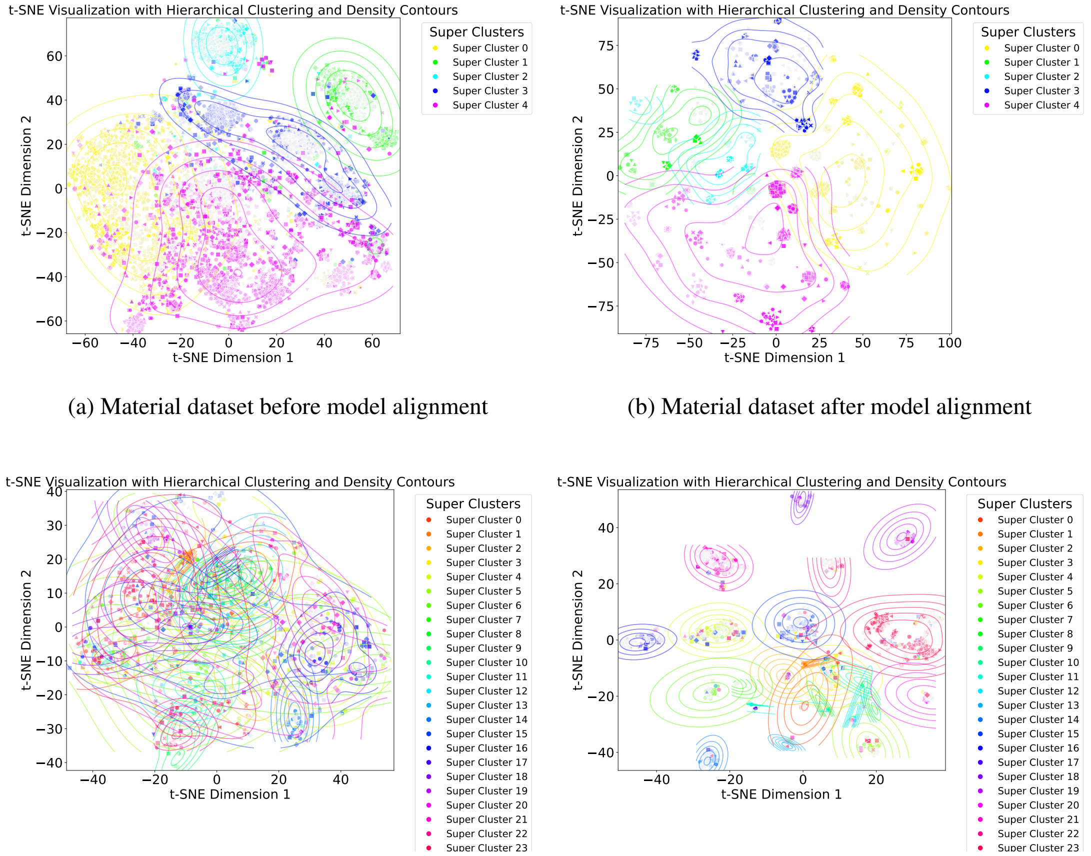
Given the specialized nature of our dataset and the requirement for hierarchical labels, finetuning is essential. Comparing our method to approaches that do not leverage hierarchical labels is inequitable, as they are inherently less effective for this task. Our approach simplifies training by eliminating HEAL loss hyperparameter tuning, unlike HiMulCon (Zhang et al., 2022), which requires extensive tuning of penalty parameters for optimal results. While HiMulCon focuses on root-level classification in vision datasets, our method aligns embeddings across all hierarchical depths. We optimize hierarchical metrics such as classification, retrieval, and hallucination indirectly through the HEAL loss, ensuring a robust alignment with the hierarchical structure.

For these reasons, we evaluate the performance of HEAL using the baseline model SciNCL, both without and with hierarchical alignment on our diverse specialized datasets. We evaluate performance using hierarchical metrics to capture nuances of hierarchical label structures in retrieval, classification, and hallucination assessments as presented in Appendix Table 2 .

4.3 Results

Table 1 summarizes the performance metrics for three datasets (Healthcare, Materials, Applied Mathematics, and Cybersecurity) across three tasks: classification, retrieval, and hallucination evaluation. The aligned model corresponds to the embedding model trained using the HEAL loss, whereas the non-aligned model corresponds to the original embedding model without HEAL-based training. The metrics are reported for both non-aligned and aligned SciNCL embeddings, demonstrating the significant impact of HEAL on improving performance. Figure 2 illustrates hierarchical embedding alignment achieved through HEAL training, resulting in well-separated super and subclusters for the Materials and Healthcare datasets which enhances the performance of downstream tasks. The density contours, computed via Kernel Density Estimation (KDE), highlight the underlying clustering structure by depicting regions of high and low embedding concentration. In subplots (a) and (c), the embeddings before model alignment appear more dispersed, indicating weaker intra-cluster cohesion and greater overlap between different data regions. However, in subplots (b) and (d), after model alignment, the contours become more compact and well-separated, signifying improved structural coherence and enhanced discriminability of the learned representations. This transformation suggests that alignment enhances the model's ability to encode meaningful relationships, ultimately improving feature organization and representation learning within the embedding space. The increased cluster compactness and separation indicate a more refined, task-specific feature space, which is crucial for downstream applications such as classification and retrieval.

First, we evaluate the performance on document-level tasks using hierarchical labels. Specifically, we assess the ability of the hierarchical classifier to predict hierarchical labels in the classification task.



(a) Material dataset before model alignment

(b) Material dataset after model alignment

(c) Healthcare dataset before model alignment

(d) Healthcare dataset after model alignment

Figure 2: Embedding visualizations for the Material and Healthcare datasets, projected using t-SNE for dimensionality reduction. The density contours represent the kernel density estimation (KDE) of the embeddings in the 2D space, highlighting the clustering structure. Subplots show the Material dataset, (a) before and (b) after model alignment, and the Healthcare dataset, (c) before and (d) after model alignment. The contours reveal changes in the density distribution of embeddings, emphasizing the effect of alignment on cluster organization and separability.

Table 1: Performance Metrics Across Datasets (Healthcare, Materials, Cyber, Applied Mathematics) for Aligned and Non-aligned Embeddings for $k = 10$

Task	***Metric**	**Healthcare**	**Materials**	**Cyber**	**Applied Mathematics**				
		Non-aligned	**Aligned**	**Non-aligned**	**Aligned**	**Non-aligned**	**Aligned**	**Non-aligned**	**Aligned**
Classification	F1 Score	0.5164	0.6588	0.6469	0.990	0.7130	0.8151	0.7541	0.8048
	Precision	0.5134	0.6590	0.6453	0.990	0.6975	0.8121	0.7415	0.8112
	Recall	0.5194	0.6586	0.6485	0.990	0.7293	0.8180	0.7672	0.7985
Retrieval	Precision@k	0.3103	0.4983	0.4787	0.9707	0.6397	0.7518	0.6576	0.7636
	Recall@k	0.0164	0.0290	0.0058	0.0116	0.0112	0.0133	0.0182	0.0212
	MRR	1.6259	2.2525	1.6541	2.9972	2.7538	3.1482	2.9065	3.2245
	nDCG@k	0.3752	0.5908	0.4982	0.990	0.6781	0.7908	0.7187	0.8280
Hallucination	FPR@k	0.9386	0.8771	0.8534	0.0878	0.7968	0.6236	0.8191	0.6529
	Severity	0.7306	0.5533	0.6041	0.0644	0.4402	0.3654	0.4119	0.3353

Additionally, we quantify the retrieval of documents from the same hierarchical category based on a query document to characterize retrieval accuracy and evaluate hallucinations. The results presented in table 1 demonstrate that HEAL significantly improves hierarchical classification metrics across all datasets. For the Healthcare dataset, the Hierarchical F1 Score improves from 0.5164 to 0.6588, reflecting a more accurate representation of hierarchical labels. Similarly, the Materials dataset achieves near perfect classification metrics (F1 Score, Precision, Recall = 0.99) with aligned embeddings, while the most challenging Healthcare dataset (4 depth cluster label) sees improvements in F1 Score from 0.5164 to 0.6588. In retrieval tasks, HEAL aligned embeddings consistently outperform non-aligned embeddings across all metrics. For the Healthcare dataset, Hierarchical MRR improves from 1.6259 to 2.2525, and nDCG@k increases from 0.3752 to 0.5908 where $k = 10$, indicating better ranking and retrieval relevance. The Materials dataset achieves a dramatic increase in retrieval precision, with Precision@k rising from 0.4787 to 0.9707, while nDCG@k reaches 0.99, showcasing near-perfect retrieval performance. For the Cyber dataset, aligned embeddings yield an MRR improvement from 2.7538 to 3.1482 and a corresponding nDCG@k increase from 0.6781 to 0.7908. Hallucination metrics further underscore the superiority of HEAL. Aligned embeddings reduce hallucination rates significantly across all datasets. For the Healthcare dataset, FPR@k drops from 0.9386 to 0.8771, and severity decreases from 0.7306 to 0.5533, indicating fewer irrelevant or misleading retrievals. The Materials dataset shows the most striking improvement, with FPR@k reduced from 0.8534 to 0.0878 and severity declining from 0.6041 to 0.0644, nearly eliminating hallucination tendencies. For the Cyber dataset, aligned embeddings lower FPR@k from 0.7968 to 0.6236 and severity from 0.4402 to 0.3654.

Next, we evaluate the performance of aligned RAG in retrieving the correct documents for generated queries to augment the LLM and minimize hallucinations. From each test dataset, we randomly sampled 100 documents and generated 10 Q&A pairs per document using the LLAMA-3.1 70B model, resulting in a total of 1,000 Q&A pairs for each dataset. Each Q&A pair was tagged with the corresponding document from which it was generated. The prompt used for Q&A generation was as follows: *"First, provide a concise summary of the following abstract that emphasizes its key concepts and hierarchical relationships. Then, based on this summary, generate 10 unique, nuanced Q&A pairs. Focus on creating questions that delve into specialized details of the hierarchical concepts discussed."* The generated queries were used to fetch documents via both aligned and unaligned models. We assessed the ability of each model to correctly retrieve the original document and evaluated the rank/order of retrieval. On average, the unaligned model achieved an MRR of 0.273 and a Recall@10 of 0.415. These metrics represent regular retrieval scores, not hierarchical scores. In contrast, the aligned model significantly improved performance, achieving an MRR of 0.514 and a Recall@10 of 0.731, demonstrating its superior ability to retrieve the correct set of documents. Furthermore, when integrating RAG with LLAMA-3.1 70B for generating answers from the queries and retrieved documents, the unaligned model produced a ROUGE score of 0.42, while the aligned model achieved a ROUGE score of 0.68. This highlights the impact of alignment on improving the quality and relevance of generated responses.

5 Conclusion

In this work, we introduced HEAL, a novel framework for aligning embeddings in RAG systems through hierarchical fuzzy clustering and matrix factorization, integrated within a contrastive learning paradigm. HEAL effectively computes level-specific contrastive losses and applies hierarchical penalties to align embeddings with domain-specific structures, enhancing both retrieval relevance and classification performance. Experimental results across diverse domains — Healthcare, Materials Science, Cybersecurity, and Applied Mathematics — demonstrate HEAL's capability to significantly improve retrieval accuracy and mitigate hallucinations in LLM-based systems. By bridging hierarchical semantics with contrastive alignment, HEAL establishes itself as a versatile and robust tool for advancing RAG methodologies, enabling more precise, reliable, and domain-adaptive applications of large language models.

6 Acknowledgement

This manuscript has been approved for unlimited release and has been assigned LA-UR-24-32672. This research was funded by the Los Alamos National Laboratory (LANL) Laboratory Directed Research and Development (LDRD) program under grants 20230067DR & 20230287ER and supported by LANL's Institutional Computing Program, and by the U.S. Department of Energy National Nuclear Security Administration under Contract No. 89233218CNA000001.

References

Luca Bertinetto, Joao F Henriques, and Philip HS Torr. 2020. Making better mistakes: Leveraging class hierarchies with deep networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1251–1259.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR.

Jia Deng, Alexander C Berg, and Li Fei-Fei. 2014. Large-scale object classification using label relation graphs. In *European Conference on Computer Vision*, pages 48–64. Springer.

Jia Deng, Sanjeev Satheesh, Alexander C Berg, and Fei-Fei Li. 2011. Hierarchical semantic indexing for large scale image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 785–792. IEEE.

Maksim E Eren, Manish Bhattarai, Robert J Joyce, Edward Raff, Charles Nicholas, and Boian S Alexandrov. 2023. Semi-supervised classification of malware families under extreme class imbalance via hierarchical non-negative matrix factorization with automatic model selection. *ACM Transactions on Privacy and Security*, 26(4):1–27.

Joshua Goodman. 2001. Classes for fast maximum entropy training. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 561–564. IEEE.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

Zi Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yanlin Xu, Etsuko Ishii, Yeonseung Bang, Andrea Madotto, and Pascale Fung. 2023. A survey of hallucination in natural language generation. *arXiv preprint arXiv:2301.07128*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Urvashi Khandelwal, Mikel Artetxe, Hailey Schoelkopf, Moin Nadeem Sung, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Fábio M Miranda, Niklas Köhnecke, and Bernhard Y Renard. 2023. Hiclass: a python library for local hierarchical classification compatible with scikit-learn. *Journal of Machine Learning Research*, 24(29):1–17.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings. In *The 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*, Abu Dhabi. Association for Computational Linguistics. 7-11 December 2022. Accepted for publication.

Nicholas Solovyev, Ryan Barron, Manish Bhattarai, Maksim E. Eren, Kim Ø. Rasmussen, and Boian S. Alexandrov. 2023. Interactive distillation of large single-topic corpora of scientific papers. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1000–1005.

Raviteja Vangara, Manish Bhattarai, Erik Skau, Gopinath Chennupati, Hristo Djidjev, Tom Tierney, James P Smith, Valentin G Stanev, and Boian S Alexandrov. 2021. Finding the number of latent topics with semantic non-negative matrix factorization. *IEEE access*, 9:117217–117231.

Shu Zhang, Ran Xu, Caiming Xiong, and Chetan Ramaiah. 2022. Use all the labels: A hierarchical multi-label contrastive learning framework.

A Appendix

A.1 Evaluation Metrics

Table 2 provides a comprehensive overview of the metrics utilized to quantify different downstream tasks such as hierarchical classification and hierarchical retrieval.

Metric	**Formula**	**Description**
Hierarchical Relevance	$\text{Relevance}(q, r) = \frac{1}{L} \sum_{l=0}^{L-1} \delta(y_q^l, y_r^l)$	Average label match across hierarchy levels
Hierarchical Precision@k	$\frac{1}{k} \sum_{i=1}^{k} \text{Relevance}(q, r_i)$	Fraction of hierarchically relevant documents among top k.
Hierarchical Recall@k	$\frac{\sum_{i=1}^{k} \text{Relevance}(q,r_i)}{\sum_{r \in \text{Relevant}(q)} \text{Relevance}(q,r)}$	Fraction of hierarchically relevant documents retrieved.
Hierarchical nDCG@k	$\frac{\sum_{i=1}^{k} \frac{2^{\text{Relevance}(q,r_i)}-1}{\log_2(i+1)}}{\sum_{i=1}^{k} \frac{2^{\text{IdealRelevance}(q,r_i)}-1}{\log_2(i+1)}}$	Discounted gain based on hierarchical relevance.
Hierarchical F1 Score	$\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$	Balance between hierarchical precision and recall.
Hierarchical Severity	$1 - \frac{\sum_{i=1}^{k} \text{Relevance}(q,r_i)}{k}$	Measures retrieval of irrelevant documents in hierarchical setting.
Hierarchical False Positive Rate@k	$\frac{\text{Irrelevant hierarchical documents in top } k}{k}$	Fraction of irrelevant hierarchical documents among top k.

Table 2: Hierarchical Metrics for classification, retrieval and hallucination