# Interpret and Control Dense Retrieval with Sparse Latent Features

**Hao Kang**
Carnegie Mellon University
Pittsburgh, PA 15213
haok@andrew.cmu.edu

**Tevin Wang**
Carnegie Mellon University
Pittsburgh, PA 15213
tevinw@andrew.cmu.edu

**Chenyan Xiong**
Carnegie Mellon University
Pittsburgh, PA 15213
cx@cs.cmu.edu

## Abstract

Dense embeddings deliver strong retrieval performance but often lack interpretability and controllability. This paper introduces a novel approach using sparse autoencoders (SAE) to interpret and control dense embeddings via the learned latent sparse features. Our key contribution is the development of a retrieval-oriented contrastive loss, which ensures the sparse latent features remain effective for retrieval tasks and thus meaningful to interpret. Experimental results demonstrate that both the learned latent sparse features and their reconstructed embeddings retain nearly the same retrieval accuracy as the original dense vectors, affirming their faithfulness. Our further examination of the sparse latent space reveals interesting features underlying the dense embeddings and we can control the retrieval behaviors via manipulating the latent sparse features, for example, prioritizing documents from specific perspectives in the retrieval results.

## 1 Introduction

In the realm of information retrieval, dense embeddings derived from large language models (LLMs) have achieved state-of-the-art performances (Khattab and Zaharia, 2020; Reimers, 2019). While these representations offer remarkable accuracy in matching queries to documents, their "black-box" nature poses challenges in applications that demand transparency and control, such as retrieval in bias-sensitive tasks, where users may need to understand the rationale behind the retrieved results and adjust the process to ensure fairness.

In contrast, in bag-of-word base sparse retrieval, each dimension is a meaningful word, allowing users to see why certain documents are retrieved, and making it intuitive for users to revise their query keywords to control the retrieval results. Interpretability and controllability are important for building trust with users and facilitate the wide adoption of search technologies (Croft et al., 2010).

In this paper, we present a novel approach that leverages sparse autoencoders (SAE) to interpret and control dense retrieval systems. Sparse autoencoders have recently been used to improve the interpretability of LLMs by transforming neuron activation patterns into sparse dictionaries (Bricken et al., 2023; Templeton et al., 2024). We upgrade this approach to dense embeddings, incorporating a retrieval-oriented recovery loss which ensures the extracted sparse features remain faithful for retrieval, forming the basis of our interpretability analysis.

Our experiments demonstrate the success of this approach. Retrieval using the learned latent sparse features and their reconstructed embeddings both recover the majority of the original dense retrieval accuracy on the MsMarco and Beir benchmarks, ensuring that these features offer genuine interpretability rather than an illusion. Then we explore the interpretability of these sparse features with Neuron to Graph (N2G) approach (Foote et al., 2023), and discover that various fine-grained concepts have been captured in the latent sparse space.

To understand controllability through latent features, we conduct quantitative studies by amplifying query-relevant features, which successfully improved retrieval accuracy on the manipulated embeddings, both on the query side and the document side. Then, we perform case studies on multi-perspective queries and confirm that selectively manipulating sparse features from a specific perspective causes the reconstructed embeddings to prioritize documents from that perspective during retrieval. Our source code and extracted features are available at GitHub [1].

## 2 Methodology

In this section, we describe the methodology used to train the sparse autoencoder with our retrieval-

[1] https://github.com/cxcscmu/embedding-scope



*Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 700–709
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

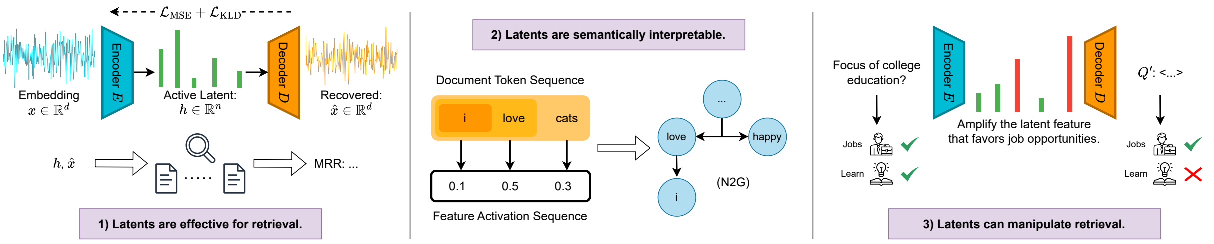



Figure 1: An overview of our framework. We first train the $k$-sparse autoencoder with our retrieval-oriented contrastive loss, which produces sparse latent features that are effective for retrieval. Next, we interpret these latents using N2G approach and demonstrate controllability via retrieval on the manipulated embeddings.

oriented recovery loss.

As illustrated in Figure 1, for an embedding vector $x \in \mathbb{R}^d$, we employ the $k$-sparse autoencoder as proposed in Makhzani and Frey (2013), which controls the number of active latent features using the TopK activation function. The encoder and decoder are described in Equation 1, where $n$ denotes the latent dimension for $W_{\text{enc}} \in \mathbb{R}^{n \times d}$. The reconstructed embedding is represented by $\hat{x} \in \mathbb{R}^d$.

$$\begin{aligned} h &= \text{TopK}(W_{\text{enc}}(x - b_{\text{dec}}) + b_{\text{enc}}) \\ \hat{x} &= W_{\text{dec}} h + b_{\text{dec}} \end{aligned} \tag{1}$$

Building on previous efforts to extract interpretable features from LLMs (Gao et al., 2024; Bricken et al., 2023; Lieberum et al., 2024), we incorporate mean-squared error (MSE) as part of the training objective for reconstruction. By minimizing the squared differences, MSE forces each dimension of the reconstructed embedding to closely approximates the original value.

However, the focus of MSE is to minimize the error for individual points in the embedding space. It does not explicitly account for the relative positioning. For information retrieval, embeddings are typically divided into queries and documents, with the need to effectively capture the relevance between a query and its associated documents.

Therefore, we employ contrastive learning via Kullback–Leibler divergence (KLD) to ensure that the distribution of reconstructed query and document embedding aligns with the original (Xiong et al., 2021; Liu et al., 2022). The formulation of the loss function is presented in Equation 2, where $q$ represents the query embedding, $D^+$ denotes the relevant documents, and $f(q, d)$ computes the retrieval score, such as dot product.

$$\begin{aligned} \mathcal{L}_{\text{KLD}} &= \sum_q \sum_{d \in D^+} P(q, d) \times \log \frac{P(q, d)}{P(\hat{q}, \hat{d})} \\ \text{where } P(q, d) &= \frac{e^{f(q,d)}}{\sum_{D^+} e^{f(q,d)}} \end{aligned} \tag{2}$$

In short, the $k$-sparse autoencoder is trained with MSE for accurate reconstruction and KLD to preserve the query-document relationship.

## 3 Experiments

This section outlines the training procedures for the $k$-sparse autoencoder and our experiments on interpretability and controllability.

**Training Procedures.** We train the autoencoder on top of the base-sized BGE model [2], which was trained on diverse tasks such as retrieval, classification, and semantic similarity (Xiao et al., 2023). Embeddings are generated from the MSMARCO dataset, containing 8.8M passages for retrieval tasks (Bajaj et al., 2016). Details of the training hyperparameters are available in Appendix A.

For evaluation, we first calculate MSE on the validation queries and their relevant documents. We then perform dense retrieval on the reconstructed embeddings and sparse dot product retrieval on the latent features. Reported metrics include mean reciprocal rank (MRR), precision at rank 10 (P@10), and recall at rank 10 (R@10).

For generalizability on diverse retrieval tasks, we additionally evalute the sparse autoencoder on datasets from the BEIR benchmark, such as TREC-COVID, NATURALQUESTIONS, and DBPEDIAENTITY (Kwiatkowski et al., 2019; Hasibi et al., 2017; Thakur et al., 2021). Additionally, we investigate the impact of the base embedding by applying our approach to an alternative embedding model, MINICPM [3] (Hu et al., 2024).

[2] https://huggingface.co/BAAI/bge-base-en-v1.5
[3] https://huggingface.co/openbmb/MiniCPM-Embedding

Table 1: Reconstruction evaluation of sparse latent features and the reconstructed embeddings learned by our $k$-sparse autoencoder from the BGE model. MSE measures the embedding differences between original and reconstructed embeddings. Results for the alternative MINICPM embedding model can be found in Appendix E.

| | MsMarco | | | | Beir | | | |
|---|---|---|---|---|---|---|---|---|
| | **MSE** | **MRR** | **P@10** | **R@10** | **MSE** | **MRR** | **P@10** | **R@10** |
| Original | – | 0.3605 | 0.0649 | 0.6211 | – | 0.3699 | 0.0891 | 0.5415 |
| Sparse Latent (K=32) | – | 0.2721 | 0.0507 | 0.4869 | – | 0.2420 | 0.0581 | 0.3590 |
| Sparse Latent (K=64) | – | 0.3062 | 0.0564 | 0.5406 | – | 0.2923 | 0.0708 | 0.4212 |
| Sparse Latent (K=128) | – | **0.3306** | **0.0601** | **0.5760** | – | **0.2981** | **0.0735** | **0.4461** |
| Reconstructed (K=32) | 0.00022 | 0.2984 | 0.0552 | 0.5291 | 0.00043 | 0.2549 | 0.0619 | 0.3768 |
| Reconstructed (K=64) | 0.00017 | 0.3194 | 0.0583 | 0.5589 | 0.00033 | 0.2913 | 0.0721 | 0.4361 |
| Reconstructed (K=128) | **0.00011** | **0.3455** | **0.0626** | **0.5991** | **0.00019** | **0.3407** | **0.0818** | **0.4954** |

**Interpretability Study.** To assess interpretability, we generate N2G explanations (Foote et al., 2023). N2G provides an automated approach to interpret the behavior of individual neurons by converting their activations into graph-based representations. It identifies the most relevant tokens that strongly activate a neuron and focuses on them by pruning the surrounding, less relevant context. This process isolates the essential patterns that contribute to the neuron's activation.

Additionally, N2G enriches the dataset by replacing key tokens with high-probability substitutes, generating variations that maintain high activation levels. By doing so, the method captures a broader and more nuanced understanding of the neuron's behavior, revealing how it responds to different inputs while maintaining its core functionality. This combination of pruning and augmentation ensures that the interpretability of each neuron is both concise and comprehensive (Foote et al., 2023).

For each feature, we create a training set of 512 samples by selecting the highest-activating documents. We then perform forward passes on prefix sequences to extract activation sequences, which are input into N2G to construct trie representations for each feature. GPT-4O-MINI is used to interpret each trie's semantic meaning.

**Controllability Study.** In the controllability experiments, we explore how amplifying sparse latent features based on relevance can influence retrieval. The experiments involve manipulating document and query embeddings.

For document manipulation, we amplify the latent feature of relevant documents in the dimension corresponding to the highest activation of each query. The modified latent features are then decoded to reconstruct the document embeddings for retrieval. For query manipulation, we amplify query features in the dimension most activated by relevant documents. A grid search determines the appropriate amplification level, starting with the smallest value of latent features at 0.0004, incremented by a factor of 2 each step.

On the other hand, we explore binary perspective queries, structured to have two distinct categories of potential document matches in our control experiments. By amplifying the latent features associated with these categories, we assess whether manipulating a particular feature leads to a greater prevalence of one category over the other during retrieval on the reconstructed embeddings.

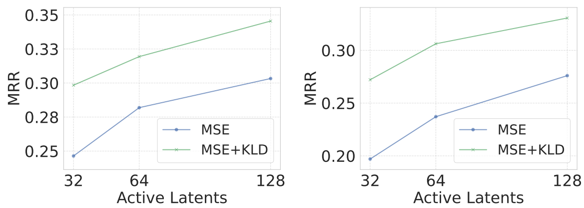
## 4 Evaluation

In this section, we present the evaluated results for each experiment in Section 3 and discuss the underlying insights that are critical for our findings.

### 4.1 Retrieval Performance

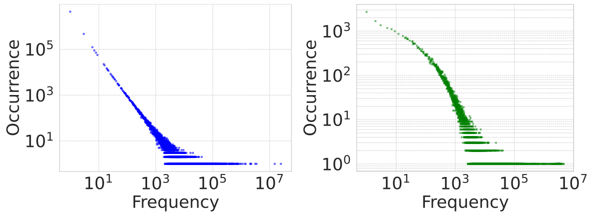
The final results in Table 1 confirm the robustness of the reconstruction. With K=128 active features in the latent space, the MSE on the MSMARCO dataset is 0.0001, and the MRR reaches 0.3455, closely aligning with the original score of 0.3605. Notably, the features extracted by the sparse autoencoder also prove valuable for retrieval, achieving an MRR of 0.3306. This utility strengthens our confidence that the interpretability analysis provides genuine insights rather than illusory interpretations.

We further assessed the impact of contrastive loss through an ablation study, comparing models trained with MSE alone against those incorporating contrastive loss. All other conditions were kept identical to ensure a fair comparison. As presented in Figure 2, the model trained with contrastive loss consistently outperforms the baseline across all latent dimensions. Notably, retrieval on sparse features improves the MRR to 0.3306, compared to 0.2760. Even though both models experience per-




(a) MSMARCO (Rec.) (b) MSMARCO (Spr.)

Figure 2: Retrieval performance of reconstructed (Rec.) embeddings and the sparse latent features (Spr.) before and after the contrastive loss KLD is applied on MSMARCO using BGE as the embedding model. Results on BEIR can be found in Appendix B.




(a) Unigram Bag-of-Words (b) Sparse Latent Features

Figure 3: Frequency distribution comparison between bag-of-words and sparse latent features in MSMARCO using BGE as the embedding model. The high-frequency region is characterized by a small number of words that occur with extreme regularity, whereas the low-frequency region consists of a large proportion of words that appear only a limited number of times throughout the dataset.

formance drop for retrieval on the BEIR dataset, models trained with contrastive loss demonstrate better resilience, suggesting stronger robustness across diverse retrieval tasks.

### 4.2 Interpretability Study

As illustrated in Figure 3, the learned sparse latent features also follow Zipf's law, but its distribution is less head-heavy. This is interesting as top-ranking features in the bag-of-words model are often common stop words, but the sparse latent features may skip these stop words and capture fine-grained and conceptually meaningful categories. Representative feature examples extracted by N2G from different segments of the distribution are provided in Table 2, while the top activated features for a sampled document in MSMARCO dataset are detailed in Table 3. Additional examples can be found in Appendix C.

### 4.3 Controllability Study

As shown in Figure 4, we observe a clear trend of improvement in both MRR and P@10 as the amplification of relevant sparse latent features increases. This demonstrates the controllability of latent features in influencing the retrieval process within the reconstructed embeddings. Specifically, as more relevance information is injected into the latent space, the retrieval scores improve. Notably, with document manipulation, the MRR reaches a peak value of 1.0 at the largest amplification level. It is also not surprising to see the performance drop on the query side when the manipulation is too strong—doubles the typically latent feature values—as it may break the reconstructed embedding.

Table 2: Examples of sparse latent features using BGE as the embedding model explained by N2G from different parts of the frequency distribution.

| **Region** | **Description from N2G** |
|---|---|
| Head | media, production, television, entertainment |
| | fashion, appearance, behavior, transformation |
| | opera, drama, music, performance, composer |
| Torso | korea, seoul, music, culture, tourism |
| | sports, injuries, protocols, regulations |
| | location, community, development, services |
| Tail | health, pain, injury, trauma, disorders |
| | growth, improvement, learning, strategy |
| | finance, investment, market, companies |

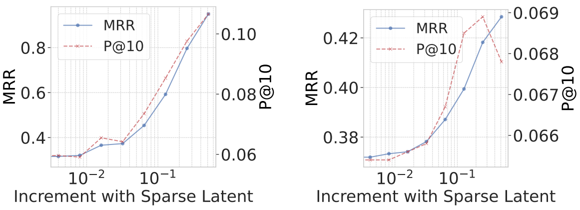
Table 3: Top activated features using BGE as the embedding model from the document "A few people reported that they paid their attorney as little as $50 per hour, and a few reported paying as much as $400 to $650 per hour. But the vast majority paid between $150 and $350 per hour, with $250 being the most commonly reported fee. The survey asked respondents about a number of things, including: 1 how much their divorce attorney charged per hour. 2 how much their divorce cost. 3 the number of issues that they resolved out of court and in court. 4 whether their spouse contested the case. 5 how long the divorce took from start to finish."

| **Description from N2G** |
|---|
| 1. cost, pricing, expenses, rates, income |
| 2. time, duration, sleep, hours, minutes |
| 3. government, law, agencies, constitution, enforcement |
| 4. tennis, courts, wimbledon, justices, decisions |
| 5. health, anxiety, symptoms, stress, concerns |

Table 4 presents one example of controlling the retrieval results by manipulating the reconstructed query embeddings via the latent space. It shows that amplifying the targeted feature dimension effectively biases the retrieval results towards the corresponding perspective, i.e., "job" (84340) or "learning" (179723). This indicates that the learned faithful latent space provides a new mechanism to

Table 4: Features for the binary perspective query "What is the primary focus of a university education?" and the top result after dense retrieval on the reconstructed embeddings using BGE as the embedding model. Feature activations were amplified by 0.5. B/A displays the number of documents related to the feature before and after the amplification on $k = 5$ retrieval.

| Feature ID | Description from N2G | Retrieved Document | B/A |
|---|---|---|---|
| 84340 | employment, salary, wages, jobs, bonuses | "...prepare people to work in various sectors of the economy or areas of culture..." | 2/3 |
| 179723 | growth, improvement, learning, strategy, development | "...for students to own knowledge, hone capacities, develop personal and social responsibility..." | 3/5 |




(a) Manipulated Document (b) Manipulated Query

Figure 4: Improvement in retrieval scores on manipulated documents and queries by amplifying relevant sparse latent features across varying amounts using BGE as the embedding model. The x-axis is in logarithmic scale for better visualizing the trends since each step gets incremented by a factor of 2.

control the retrieval behavior which leads to many potential applications, for example, in enhancing the safety with human intervention in dense retrieval systems. Additional examples can be found in Appendix D.

## 5 Conclusion

In this paper, we presented a novel method that applies sparse autoencoder to enhance the interpretability and controllability of dense embedding spaces in information retrieval. Our approach, which utilizes a retrieval-oriented contrastive loss function, ensures that the sparse features extracted remain faithful for interpretation. The experimental results demonstrate that our reconstructed embeddings maintain competitive retrieval accuracy, with sparse latent features proving to be both interpretable and controllably influential on retrieval outcomes. By enabling explicit manipulation of these sparse features, we provide a means to directly influence retrieval behaviors, offering a significant advantage for applications requiring transparent and adjustable retrieval mechanisms.

## 6 Limitations

One limitation of this work is the potential for scaling. While the method demonstrates effectiveness, its scalability to larger embedding space remains to be explored. Additionally, although the sparse latent features offer strong evidence of interpretability and controllability, the relationship between these features and retrieval outcomes is still correlational, rather than causal. Thus, there is no guarantee that manipulating these features will always lead to the desired retrieval behavior. Lastly, while the sparse latent space approximates the performance of dense embeddings, it has not fully recovered the original retrieval performance, indicating room for further improvement.

## References


Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

W Bruce Croft, Donald Metzler, and Trevor Strohman. 2010. *Search engines: Information retrieval in practice*, volume 520. Addison-Wesley Reading.

Alex Foote, Neel Nanda, Esben Kran, Ioannis Konstas, Shay Cohen, and Fazl Barez. 2023. Neuron to graph: Interpreting language model neurons at scale. *arXiv preprint arXiv:2305.19911*.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*.

Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. 2017. Dbpedia-entity v2: a test collection for entity search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1265–1268.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Diederik P Kingma. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *arXiv preprint arXiv:2408.05147*.

Zhenghao Liu, Han Zhang, Chenyan Xiong, Zhiyuan Liu, Yu Gu, and Xiaohua Li. 2022. Dimension reduction for efficient dense retrieval via conditional autoencoder. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5692–5698, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Alireza Makhzani and Brendan Frey. 2013. K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*.

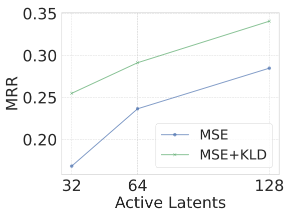
N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

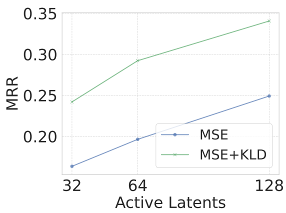
Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2023. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*.

Wenhan Xiong, Xiang Lorraine Li, Srinivasan Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Wen-tau Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oğuz. 2021. Answering complex open-domain questions with multi-hop dense retrieval. *International Conference on Learning Representations*.




(a) BEIR (Rec.)




(b) BEIR (Spr.)

Figure 5: Retrieval performance of reconstructed (Rec.) embeddings and the sparse latent features (Spr.) before and after the contrastive loss KLD is applied on BEIR using BGE as the embedding model.

# A Training Procedures

During training, we employ the Adam optimizer (Kingma, 2014) with a batch size of 512 across 128 total epochs. The initial learning rate is set to $1 \times 10^{-3}$ and is progressively reduced using the cosine annealing scheduler (Loshchilov and Hutter, 2016). We sample 16 relevant documents per query from the original embedding space to compute the loss function in an efficient manner.

# B Ablation Study

This section presents the ablation study, comparing models trained with MSE alone against those incorporating contrastive loss on the BEIR dataset. The comparison is illustrated with Figure 5.

# C Interpretability Study

In our interpretability analysis, we utilize the N2G approach to interpret latent features extracted by the autoencoder. Sampled features from different parts of the frequency distribution (i.e. head, torso, tail) are shown in Table 6 along with their N2G explanations. Activated features and their associated semantic concepts for a subset of queries from MSMARCO dataset are displayed in Table 7.

# D Controllability Study

This section examines how feature activations can control retrieval on binary perspective queries. Tables 4 presents how feature amplification affects the number of relevant documents retrieved before and after (B/A) over the binary perspective queries "What is a key factor in the spread of infectious diseases?" and "What is a major influence on automotive emissions?".

# E Role of Base Embedding

This section explores the transferability of our method across different embedding models. As illustrated in Table 8, our approach demonstrates consistent performance when applied to the MINICPM embedding. However, we observe a noticeable decline in retrieval accuracy when using the sparse autoencoder with $K = 32$ active features. This reduction may be attributed to the significantly larger embedding dimension involved, which is three times the size of BGEBASE. This increased dimensionality likely necessitates a greater number of active features to support the retrieval task. Additionally, the results of our interpretability analysis and controllability study, conducted using the MINICPM embedding, are presented in Tables 9, 10, and 11.

Table 5: Manipulation over for the binary perspective queries "What is a key factor in the spread of infectious diseases?" and "What is a major influence on automotive emissions?" by amplifying the perspective latent features using BGE as the embedding model.

| Feature ID | Description from N2G | Retrieved Document | B/A |
|---|---|---|---|
| 15678 | health, nutrition, immune, disease, metabolism | "...1 Route of entry of the pathogen and the access to host regions that it gains. 2 Intrinsic virulence of the particular organism..." | 2/3 |
| 53246 | demographics, migration, populations, countries, socioeconomic | "...Learn how our modern way of life contributes to the spread and emergence of disease. 1 Globalization. 2 Climate Change. 3 Ecosystem Disturbances. 4 Poverty, Migration & War..." | 1/4 |
| 142071 | climate, weather, precipitation, seasons, diversity | "... Major smog occurrences often are linked to heavy motor vehicle traffic, high temperatures, sunshine, and calm winds...." | 2/5 |
| 155875 | automotive, engineering, mechanics, combustion, manufacturing | "...1 Driving and atmospheric conditions. 2 Mileage. 3 Vehicle age. Type of spark plug electrode 1 material. Poor vehicle maintenance. Poor quality 1 fuel. Damaged or worn sensors. Dry-rotted or cracked vacuum hoses...." | 3/5 |

Table 6: Sparse latent features from the frequency distribution using BGE as the embedding model.

| Region | Feature ID | Description from N2G |
|---|---|---|
| Head | 3 | media, production, television, entertainment |
| | 24 | fashion, appearance, behavior, transformation |
| | 30 | opera, drama, music, performance, composer |
| | 58 | health, dignity, history, identity, inquiry |
| | 82 | festival, country, music, education, rural |
| | 86 | identity, culture, lifestyle, expression, community |
| Torso | 28840 | korea, seoul, music, culture, tourism |
| | 53784 | sports, injuries, protocols, regulations |
| | 73817 | location, community, development, services |
| | 91052 | meaning, significance, language, culture |
| | 99785 | age, death, health, statistics, history |
| | 194488 | weather, precipitation, climate, population |
| Tail | 136995 | health, pain, injury, trauma, disorders |
| | 179723 | growth, improvement, learning, strategy |
| | 182171 | finance, investment, market, companies |
| | 137124 | healthcare, assessment, professionals |
| | 143764 | health, anatomy, surgery, body, women |
| | 189083 | temperature, climate, weather, humidity |

Table 7: Top activated features from a subset of queries in MsMarco dataset using BGE as the embedding model.

| Query Text | Feature ID | Description from N2G |
|---|---|---|
| "what is prism in eyeglasses" | 3125 | pattern, structure, variation, sequence |
| | 39670 | cosmetics, color, skin, makeup, stain |
| | 39122 | stimuli, patterns, response, signals, activation |
| | 114454 | Beauty, identity, color, fashion, expression |
| | 15678 | health, nutrition, immune, disease, metabolism |
| "what are the characteristics of the eucalyptus" | 14689 | pets, veterinary, animals, dog, care |
| | 15678 | health, nutrition, immune, disease, metabolism |
| | 39122 | stimuli, patterns, response, signals, activation |
| | 142071 | climate, weather, precipitation, seasons |
| | 189083 | temperature, climate, humidity, weather |
| "best wr in nfl history" | 69658 | wildcard, subsequences, activation, neuron |
| | 71882 | baseball, athletes, performance, statistics |
| | 78287 | classification, types, examples, varieties |
| | 100445 | tennis, courts, justices, championships |
| | 155393 | celebrity, entertainment, personality, humor |
| "how long is cough in children lasting" | 15678 | health, nutrition, immune, disease, metabolism |
| | 39122 | stimuli, patterns, response, signals, activation |
| | 45139 | time, duration, sleep, hours, minutes |
| | 56299 | measurements, values, dimensions, statistics |
| | 185691 | weather, forecast, conditions, cold, outlook |

Table 8: Reconstruction evaluation of sparse latent features and the reconstructed embeddings learned by our $k$-sparse autoencoder from MINICPM embedding model.

| | MsMarco | | | |
|---|---|---|---|---|
| | **MSE** | **MRR** | **P@10** | **R@10** |
| Original | – | 0.3770 | 0.0682 | 0.6519 |
| Sparse Latent (K=32) | – | 0.1908 | 0.0389 | 0.3745 |
| Sparse Latent (K=64) | – | 0.2594 | 0.0507 | 0.4870 |
| Sparse Latent (K=128) | – | **0.2953** | **0.0565** | **0.5416** |
| Reconstructed (K=32) | 0.00014 | 0.3128 | 0.0587 | 0.5613 |
| Reconstructed (K=64) | 0.00011 | 0.3397 | 0.0630 | 0.6025 |
| Reconstructed (K=128) | **0.00009** | **0.3535** | **0.0649** | **0.6207** |

Table 9: Manipulation over for the binary perspective queries “"What determines the success of rehabilitation therapy?” and “What shapes consumer decisions when buying eyewear?” by amplifying the perspective latent features using MINICPM as the embedding model.

| **Feature ID** | **Description from N2G** | **Retrieved Document** | **B/A** |
|---|---|---|---|
| 183 | energy, transformation, healing, vitality, balance | “...Setting goals is the best way to achieve a successful rehabilitation outcome....” | 0/0 |
| 4857 | time, duration, intervals, periods, estimation | “With treatment, a few people recover in a year or less. For the vast majority, though, treatment and the recovery process take three to seven years, and in some cases even longer.” | 0/5 |
| 39423 | health, vision, care, eye, conditions | “What time of the day to have eye exam to get prescription eye glasses? I need a new pair of glasses (near sighted + other). I wonder it makes a little difference to go in the morning or afternoon or evening. I wonder if the eyesight is better in the morning after a night’s sleep? Should I get eye exam when the eyesight is in best or worst condition?” | 1/5 |
| 161546 | glasses, eyewear, sunglasses, styles, features | “When buying eyeglasses, the frame you choose is important to both your appearance and your comfort when wearing glasses. But the eyeglass lenses you choose influence four factors: appearance, comfort, vision and safety.” | 2/4 |

Table 10: Sparse latent features from the frequency distribution using MINICPM as the embedding model.

| **Region** | **Feature ID** | **Description from N2G** |
|---|---|---|
| Head | 25 | health, medical, conditions, females, diagnosis |
| | 97 | patterns, sequences, triggers, signals, behavior |
| | 183 | energy, transformation, healing, vitality, balance |
| | 197 | signals, patterns, thresholds, responses, stimuli |
| | 207 | television, advertising, marketing, entertainment, engagement |
| | 236 | ot, Rep, neuron, activation, subsequence |
| Torso | 146050 | trading, hours, market, business, activities |
| | 188194 | Health, recreation, arts, fitness, therapy |
| | 140841 | health, wellness, community, education, environment |
| | 109917 | health, wellness, nutrition, activities, rituals |
| | 153312 | movie, technology, vehicle, animal, mechanics |
| | 154625 | analysis, patterns, activation, signals, behavior |
| Tail | 114226 | communication, education, resources, technology, collaboration |
| | 107220 | health, wellness, genetics, lifestyle, information |
| | 125167 | blood, language, difference, country, education |
| | 144165 | cellular, biological, procedures, structures, metabolism |
| | 193906 | neurobiology, stimuli, patterns, activation, response |
| | 125701 | communication, processes, information, interactions, connections |

Table 11: Top activated features from a subset of queries in MsMarco dataset using MiniCPM as the embedding model.

| Query Text | Feature ID | Description from N2G |
|---|---|---|
| "what is prism in eyeglasses" | 161546 | glasses, eyewear, sunglasses, styles, features |
| | 26168 | structure, geometry, prism, dimensions, properties |
| | 39423 | health, vision, care, eye, conditions |
| | 179744 | activation, patterns, sequences, neuron, inputs |
| | 109256 | education, activities, science, culture, resources |
| "what are the characteristics of the eucalyptus" | 47108 | neuron, activation, patterns, sequences, stimulation |
| | 56389 | characteristics, organisms, life, description, taxonomic |
| | 143997 | characteristics, features, descriptions, attributes, traits |
| | 84508 | forest, trees, timber, ecology, sustainability |
| | 134883 | Australia, Australians, territories, states, constitution |
| "best wr in nfl history" | 16624 | football, NFL, teams, players, games |
| | 179906 | receiver, wide, receptions, football, targets |
| | 147634 | history, culture, documentation, information, analysis |
| | 189070 | health, disease, communication, identity, experience |
| | 143889 | patterns, sequences, neural, interactions, responses |
| "how long is cough in children lasting" | 103545 | cough, symptoms, conditions, medical, causes |
| | 29915 | children, pediatric, development, therapy, care |
| | 174114 | lungs, breathing, pulmonary, respiratory, health |
| | 4857 | time, duration, intervals, periods, estimation |
| | 113082 | cough, chronic, symptoms, causes, prevalence |