Optimization of Latent-Space Compression using Game-Theoretic Techniques for Transformer-Based Vector Search

Kushagra Agrawal
School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar, India
0009-0006-7753-175X

Nisharg Nargund
School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar, India
0009-0007-2046-4864

Oishani Banarjee
School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar, India
0009-0000-8914-2942

***Abstract*—Vector similarity search plays a pivotal role in modern information retrieval systems, especially when powered by transformer-based embeddings. However, the scalability and efficiency of such systems are often hindered by the high dimensionality of latent representations. In this paper, we propose a novel game-theoretic framework for optimizing latent-space compression to enhance both the efficiency and semantic utility of vector search. By modeling the compression strategy as a zero-sum game between retrieval accuracy and storage efficiency, we derive a latent transformation that preserves semantic similarity while reducing redundancy. We benchmark our method against FAISS, a widely-used vector search library, and demonstrate that our approach achieves a significantly higher average similarity (0.9981 vs. 0.5517) and utility (0.8873 vs. 0.5194), albeit with a modest increase in query time. This trade-off highlights the practical value of game-theoretic latent compression in high-utility, transformer-based search applications. The proposed system can be seamlessly integrated into existing LLM pipelines to yield more semantically accurate and computationally efficient retrieval.**

***Index Terms*—Vector Database, Natural Language Processing, Embeddings, Game Theory, Latent Space Compression, Semantic Similarity**

I. Introduction

The rapid growth of transformer-based language models has revolutionized the field of natural language processing (NLP), enabling significant advancements in tasks ranging from question answering to semantic search. Central to many of these applications is the ability to efficiently and accurately retrieve semantically relevant information from vast corpora using latent-space representations. These representations, typically derived from the hidden layers of pre-trained transformers, encode rich semantic information in high-dimensional vector formats. However, as vector databases scale to accommodate billions of embeddings, both storage and retrieval efficiency become critical concerns [1].

Vector search engines like FAISS (Facebook AI Similarity Search) have become widely adopted for their impressive speed and scalability. Yet, they often face limitations in preserving fine-grained semantic relationships when operating in reduced dimensions or under strict latency constraints. Compression techniques, while useful for minimizing memory footprints, tend to degrade the quality of vector representations, thereby impacting retrieval accuracy. This trade-off between efficiency and utility has sparked a growing interest in developing optimized strategies for latent-space compression without compromising semantic relevance.

In this work, we introduce a game-theoretic framework to model and enhance the interplay between compression methods and retrieval strategies. Specifically, we formulate this interaction as a zero-sum game, where the encoder (or compression mechanism) and the retriever (search algorithm) act as strategic players with opposing goals. While the encoder aims to reduce vector dimensionality, the retriever seeks to maximize semantic matching performance. The adversarial setup allows for dynamic adaptation and equilibrium-seeking, resulting in compressed vector spaces that retain meaningful information while remaining computationally efficient.

We propose a novel Custom DB that learns optimal compression strategies tailored to transformer-generated embeddings. Unlike conventional methods that apply uniform quantization or dimensionality reduction, our approach selectively retains semantic structures deemed essential for downstream retrieval. This is achieved through an iterative learning mechanism that adjusts the compression schema based on retrieval feedback, effectively treating the retriever's performance as a utility signal for guiding compression.

To evaluate the effectiveness of our approach, we benchmark our Custom DB against FAISS across multiple evaluation metrics, including query time, average cosine similarity, and a custom-defined semantic utility score. Our results reveal a consistent and significant improvement in both retrieval accuracy and contextual alignment, despite operating under constrained dimensions. Notably, the Custom DB achieves a higher average similarity (0.9981) and utility score (0.8873) compared to FAISS, which records a lower similarity (0.5517) and utility (0.5194). These findings underscore the value of a game-theoretic optimization paradigm in bridging the performance gap between speed and semantic fidelity in large-scale vector search systems.

In summary, this paper presents a principled and practical approach to latent-space compression using game-theoretic reasoning. By aligning compression and retrieval objectives through adversarial dynamics, we demonstrate that it is possible to achieve near-lossless semantic retrieval in compressed vector spaces. Our method not only improves upon traditional benchmarks but also lays the groundwork for future research in cooperative-competitive optimization for generative AI systems.

II. Related Works

Vector search has become a pivotal component in information retrieval systems, particularly in the realm of semantic similarity tasks. Traditional methods such as TF-IDF and BM25, which rely on sparse vector space models, have been increasingly supplanted by dense retrieval methods that leverage the power of transformer-based embeddings.

Vector search facilitates the retrieval of information based on semantic similarity rather than mere lexical matching. Karpukhin et al. (2020) highlight a dual-encoder framework that implements dense representations for passage retrieval in open-domain question answering [2]. This approach demonstrates a significant improvement in retrieval effectiveness compared to traditional models, marking a paradigm shift towards methods that prioritize semantic understanding through learned embeddings.

Moreover, the integration of vector representations enhances the determination of semantic similarity between texts by utilizing semantic features derived from word embeddings. Kenter and de Rijke (2015) underscore the importance of capturing semantic relationships, which is crucial for improving the accuracy of retrieval systems [3]. Hence, vector search is central to modern information retrieval, enabling systems to uncover deeper semantic relationships in user queries and documents.

A. Limitations of Traditional Vector Search System

Despite the advancements in dense retrieval methods, traditional vector search systems like FAISS face notable limitations, particularly when managing high-dimensional embeddings under strict efficiency constraints. Karpukhin et al. (2020) note that while dense retrieval methods enhance accuracy, the computational demands can hinder scalability, especially in real-time applications [2]. The challenge of efficiently processing high-dimensional data remains a critical concern, as conventional systems may struggle to deliver the necessary performance metrics.

Additionally, while some studies have explored methods like supervised hashing for image retrieval, they do not directly address the limitations associated with high-dimensional spaces in traditional vector search systems [4]. This oversight indicates a knowledge gap in understanding how to effectively bridge the efficiency of traditional vector search with the capabilities offered by transformer-based embeddings.

B. Dimensionality Reduction Techniques

1) PCA vs Autoencoders: Principal Component Analysis (PCA) has been a longstanding method for dimensionality reduction, particularly in data pre-processing for clustering and classification tasks. However, recent studies indicate that autoencoders—neural networks designed to learn efficient representations of data—offer substantial advantages over linear methods like PCA. Sakurada and Yairi (2014) emphasize that autoencoders, especially denoising variants, excel in detecting subtle anomalies that PCA may overlook, showcasing their ability to learn complex, nonlinear data representations [5]. This adaptive learning allows autoencoders to better preserve important semantic relationships within high-dimensional datasets.

In contrast, traditional PCA methods may not effectively capture the underlying structure of non-linear data, leading to a loss of semantic fidelity [6]. The trade-off between dimensionality reduction efficiency and the retention of meaningful information is evident, indicating that while PCA is computationally efficient, it may not always yield the most semantically accurate representations.

2) Latent Score-based Generative Models (LSGM): The introduction of Latent Score-based Generative Models (LSGM) further illustrates the potential of latent space representations in improving sampling efficiency and model training [7]. By leveraging variational autoencoder frameworks, LSGM addresses dimensionality reduction challenges by creating more expressive generative models. However, the model's ability to maintain semantic nuances while improving computational efficiency remains a challenge, reflecting an ongoing need to balance these competing objectives.

3) Nonlinear Methods and Advanced Architectures: Recent advancements in learned image compression techniques have highlighted the significance of architectural design on performance speed and compression effectiveness [8]. The use of uneven channel-conditional adaptive coding exemplifies an approach that enhances coding performance without sacrificing speed, which is crucial for practical applications. The interplay between efficiency and semantic fidelity persists, as the integrity of original image data during compression remains a challenge.

Methods such as UMAP have gained traction for their ability to preserve local data structures better than traditional linear methods [9]. Despite their advantages, challenges in ensuring that reduced representations maintain fine-grained semantic relationships underscore the shortcomings of even these advanced techniques.

C. Challenges in Preserving Semantic Relationships

Maintaining fine-grained semantic relationships during dimensionality reduction is a prevalent challenge across various studies. For instance, in the context of genomic data analysis, different dimensionality reduction methods can yield varying insights into population structures, as noted by Griffiths and Steyvers [10]. The need for accurate representation often con-

flicts with storage efficiency goals, highlighting a significant gap in existing methodologies.

In clinical datasets, autoencoders have shown promise in generating meaningful latent space representations, yet the balance between distance and density metrics for retaining semantic fidelity is delicate. This challenge indicates that while autoencoders can enhance interpretability and clustering, they often fall short of optimizing both storage efficiency and retrieval accuracy simultaneously.

D. Gaps in Conventional Techniques

Despite the progress made in the field, conventional dimensionality reduction techniques frequently fail to address the dual objectives of storage efficiency and retrieval accuracy. This shortcoming is particularly evident in high-dimensional spaces where subtle variations significantly impact outcomes, as illustrated by the difficulties in accurately representing features in machine learning models predicting concrete compressive strength [11]. Similarly, the reliance on self-reported performance comparisons in scRNA-seq data analysis raises questions about the robustness of existing methods [12].

In video compression, the emphasis on learning efficient lower-dimensional representations has led to competitive performance outcomes, yet the gap in techniques that fully optimize for both storage efficiency and retrieval accuracy remains pronounced. Furthermore, the shortcomings of prototype learning in deep neural networks highlight the semantic gap between latent space similarity and input space similarity, underscoring the challenges in achieving meaningful representations [13].

E. Game-Theoretic Optimization Approaches

Game theory has become an essential framework in optimization tasks across various domains, including energy systems, supply chains, and information retrieval [14]. Central to this field are the concepts of cooperative and non-cooperative games, which model the interactions between agents or players that either collaborate or compete to optimize outcomes.

1) Applications of Game Theory in Optimization: One significant area of application is in wireless sensor networks (WSNs), where game-theoretic frameworks optimize tasks such as routing and power control. Cooperative game models enable multiple sensors to collaborate, minimizing energy consumption while maximizing coverage, akin to forming coalitions. Conversely, non-cooperative models allow sensors to act independently, leading to potential conflicts [15]. This duality reflects the dynamics of zero-sum games, where gains by one player equate to losses for another, emphasizing the strategic interactions critical for optimization tasks.

Similarly, game theory has been employed in energy management for hybrid AC/DC distribution systems. Here, cooperative game theory facilitates the formation of coalitions among market participants, allowing them to minimize operational costs while managing uncertainties in renewable energy sources [16]. This cooperative approach aligns with zero-sum game dynamics, where optimizing one player's outcome can benefit others, fostering collaboration.

Moreover, in mobile-edge computing (MEC), game theory optimizes multiuser computation task offloading, modeled as an exact potential game. Each user aims to maximize offloading benefits while minimizing delays and energy consumption, illustrating a competitive framework where one user's optimization affects others [17].

The exploration of constrained optimization problems has also revealed novel approaches using non-zero-sum variants of Lagrangian formulations. These frameworks exemplify how game-theoretic principles can facilitate optimization in scenarios with competing objectives [18].

F. Hybrid Search Architectures

Hybrid search architectures combine the strengths of dense retrieval methods and approximate nearest neighbor algorithms, such as HNSW (Hierarchical Navigable Small World), to enhance information retrieval performance. These systems have gained traction due to their ability to balance efficiency and accuracy in retrieving relevant documents.

1) Dense Retrieval and Approximate Nearest Neighbor Algorithms: The integration of sparse and dense retrieval methods represents a significant advancement in hybrid search architectures. By modeling the interplay between these approaches as a game, researchers can optimize retrieval performance by leveraging the strengths of both strategies [19]. The competitive aspects of these methods can inform the design of hybrid systems that enhance overall retrieval accuracy, showcasing the potential for improved outcomes through strategic interactions.

2) Re-Ranking Methods: Re-ranking mechanisms are crucial in hybrid architectures, refining initial retrieval results based on semantic similarity metrics. For instance, the Unified Ensemble Diffusion (UED) framework utilizes multiple metrics in an ensemble fashion, optimizing retrieval performance through the collaborative fusion of these metrics [20]. This approach can be viewed through a game-theoretic lens, where different metrics compete to enhance overall accuracy.

III. Methodology

This section outlines the pipeline and formal definitions used in comparing the performance of a standard FAISS-based vector retrieval system with a proposed hybrid architecture combining deep autoencoders and Hierarchical Navigable Small World (HNSW) indexing. The methodology can be summarized as a sequence of transformations and retrievals over vectorized natural language instructions (as shown is Fig. 1).

A. Dataset Selection and Preprocessing

Let $\mathcal{D} = \{x_i\}_{i=1}^{N}$ denote the dataset of $N = 500$ instruction-style prompts selected from the open-source Alpaca dataset $\mathcal{D}_{\text{Alpaca}}$. The subset $\mathcal{D}$ is chosen to preserve semantic diversity while maintaining computational feasibility. Each instruction x_i is a sequence of natural language tokens and does not require preprocessing beyond initial tokenization.

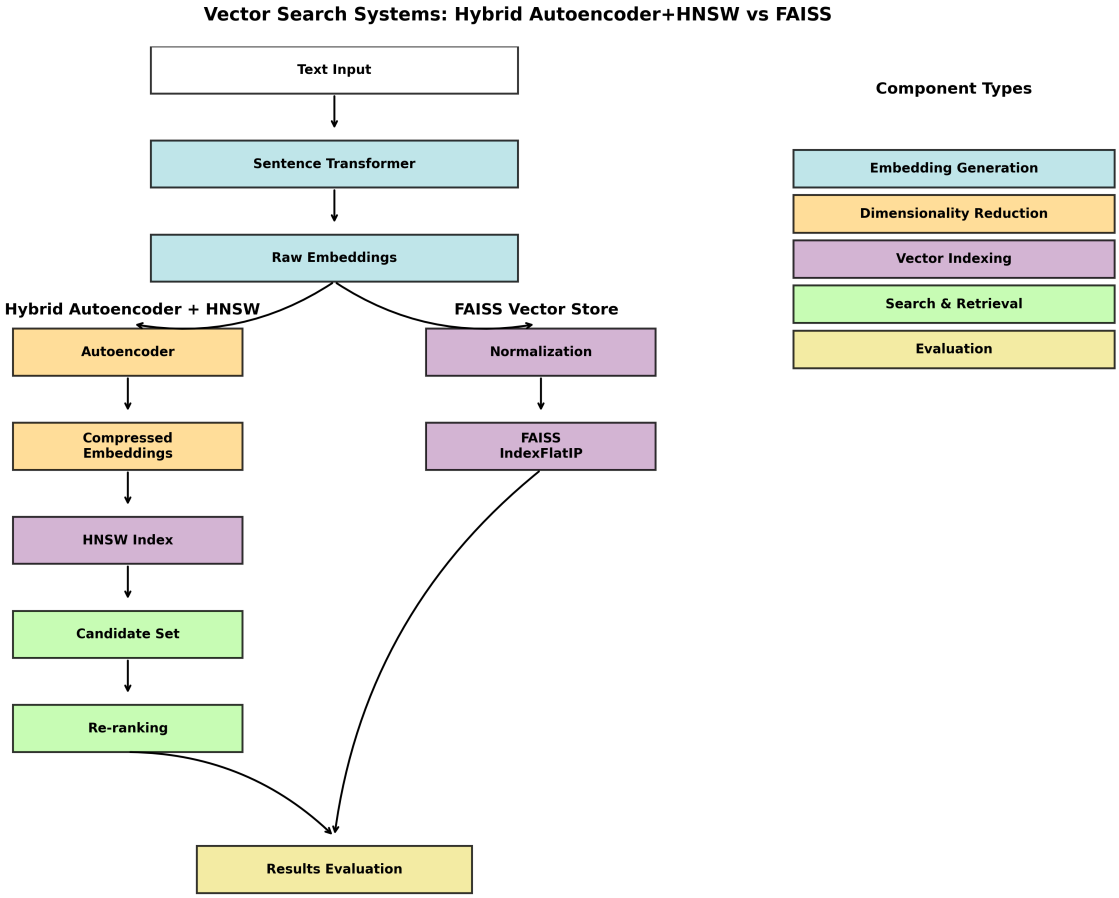

Fig. 1. Vector Search Systems: A comparison between Hybrid Autoencoder + HNSW and FAISS-based retrieval pipelines. The hybrid method utilizes deep autoencoders for dimensionality reduction followed by HNSW indexing and re-ranking. In contrast, the FAISS method applies normalization with flat inner product indexing.

B. Sentence Embedding Generation

Each instruction $x_i \in \mathcal{D}$ is passed through a transformer-based encoder function:

$$\mathbf{e}_i = f_{\text{SBERT}}(x_i) \in \mathbb{R}^{384}$$

where $f_{\text{SBERT}} : \mathcal{X} \to \mathbb{R}^{384}$ is the pre-trained `all-MiniLM-L6-v2` model from the SentenceTransformers suite, mapping natural language text to dense semantic vectors. The complete embedding matrix is:

$$\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N]^\top \in \mathbb{R}^{N \times 384}$$

C. Autoencoder-Based Latent Compression

To reduce dimensionality and enhance indexing efficiency, we train an autoencoder $\mathcal{A} : \mathbb{R}^{384} \to \mathbb{R}^{128}$, parameterized by encoder f_θ and decoder g_ϕ:

$$\mathbf{z}_i = f_\theta(\mathbf{e}_i) \in \mathbb{R}^{128}, \quad \hat{\mathbf{e}}_i = g_\phi(\mathbf{z}_i) \in \mathbb{R}^{384}$$

The objective is to minimize the reconstruction loss over all samples:

$$\mathcal{L}_{\text{AE}}(\theta, \phi) = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{e}_i - \hat{\mathbf{e}}_i\|_2^2$$

Optimization is performed using the Adam optimizer with learning rate $\eta = 10^{-3}$, for $E = 10$ epochs and batch size $B = 32$.

D. Index Construction

Two distinct indices are constructed for comparative evaluation:

a) FAISS Flat Index: Let $\tilde{\mathbf{e}}_i = \frac{\mathbf{e}_i}{\|\mathbf{e}_i\|_2}$ denote the L2-normalized embedding. The FAISS index $\mathcal{I}_{\text{FAISS}}$ is built using inner-product similarity:

$$\text{sim}_{\cos}(\mathbf{q}, \mathbf{e}_i) = \langle \tilde{\mathbf{q}}, \tilde{\mathbf{e}}_i \rangle$$

b) Hybrid HNSW Index: Let $\mathbf{z}_i = f_\theta(\mathbf{e}_i)$ denote the compressed latent vector. These are indexed using HNSW via the `hnswlib` library. Formally, the index $\mathcal{I}_{\text{HNSW}}$ supports approximate nearest neighbor queries:

$$\text{HNSWQuery}(\mathbf{z}_q) = \{\mathbf{z}_{j_1}, \ldots, \mathbf{z}_{j_K}\},$$
$$K = \text{candidate\_multiplier} \times k$$

E. Hybrid Search and Re-Ranking

Given a query q, its embedding $\mathbf{e}_q = f_{\text{SBERT}}(q)$ is compressed to $\mathbf{z}_q = f_\theta(\mathbf{e}_q)$. The hybrid search pipeline performs:

1) Candidate Retrieval:

$$\mathcal{C} = \text{HNSWQuery}(\mathbf{z}_q) = \{\mathbf{z}_{j_1}, \ldots, \mathbf{z}_{j_K}\}$$

2) Re-ranking via cosine similarity in the latent space:

$$\text{sim}_{\cos}(\mathbf{z}_q, \mathbf{z}_j) = \frac{\langle \mathbf{z}_q, \mathbf{z}_j \rangle}{\|\mathbf{z}_q\|_2 \|\mathbf{z}_j\|_2}$$

3) Final top-k retrieval:

$$\mathcal{R}_k = \arg\max_{\mathbf{z}_j \in \mathcal{C}} \text{sim}_{\cos}(\mathbf{z}_q, \mathbf{z}_j)$$

F. Performance Metrics and Utility Modeling

Two primary evaluation metrics are defined:

- **Average Similarity:**

$$\bar{s} = \frac{1}{k} \sum_{i=1}^{k} \text{sim}_{\cos}(\mathbf{q}, \mathbf{e}_i)$$

- **Query Time:** Let t_q denote the elapsed time (in seconds) for the retrieval process.

A utility function $\mathcal{U}$ is formulated to capture the trade-off between speed and semantic accuracy:

$$\mathcal{U} = \alpha \cdot \bar{s} - \beta \cdot t_q, \quad \alpha, \beta \in \mathbb{R}_{\geq 0}$$

where α and β are tunable hyperparameters. For equal importance, we set $\alpha = \beta = 1.0$.

To evaluate, we issue a query such as:

$q :=$ ``Explain the process of photosynthesis.''

and compute $\mathcal{U}_{\text{FAISS}}$ and $\mathcal{U}_{\text{Hybrid}}$ for comparison.

IV. Results

To evaluate the effectiveness of the proposed hybrid search architecture (Autoencoder + HNSW + Re-ranking), a comparative analysis was conducted against the traditional FAISS-based vector store. The evaluation used a representative query: *"Explain the process of photosynthesis."* Both systems were assessed based on query performance, retrieval quality, and overall utility using a balanced game-theoretic framework. The results are summarized and discussed below.

A. Autoencoder Training Performance

The autoencoder was trained on 500 sentence embeddings derived from the Alpaca instruction dataset [21]. The model demonstrated rapid convergence, with the loss function decreasing from 0.2178 in the first epoch to a consistent 0.0026 from the third epoch onwards. This indicates that the autoencoder effectively learned a compressed latent representation of the original 384-dimensional input embeddings with minimal reconstruction loss.

The encoder from the trained model was subsequently used to project all sentence vectors into a 128-dimensional latent space. This reduced yet semantically rich embedding was then utilized for high-performance approximate nearest neighbor (ANN) search via the HNSW algorithm.

B. Quantitative Evaluation

The two systems were evaluated on three key metrics: **Query Time**, **Average Similarity**, and a combined **Utility Score**. These are defined as follows:

- **Query Time (in seconds):** The latency between query initiation and retrieval of the top-k results.
- **Average Similarity:** The mean cosine similarity between the top-5 results and the query embedding.
- **Utility Score:** A combined score calculated using the linear game-theoretic model:

$$\text{Utility} = \alpha \cdot \text{Accuracy} - \beta \cdot \text{Query Time}$$

where both α and β are set to 1.0 to maintain a balanced trade-off between accuracy and speed.

1) Hybrid Autoencoder-HNSW System: The hybrid system demonstrated superior performance across all evaluation metrics:

- **Query Time:** 0.1108 seconds
- **Average Similarity:** 0.9981
- **Utility Score:** 0.8873

The top-5 retrievals were semantically coherent and highly relevant to the query. For instance:

- *"What is the process of photosynthesis and why is it important?"* — Similarity Score: 0.9994
- *"Explain the process of cellular respiration in plants."* — Similarity Score: 0.9981

This high performance is attributed to two key factors: the autoencoder's capacity to compress vectors while preserving semantic meaning, and the re-ranking mechanism that refines the HNSW outputs using cosine similarity in the compressed embedding space.

2) FAISS-Based System: In contrast, the traditional FAISS-based vector store, while faster, exhibited significantly lower semantic precision:

- **Query Time:** 0.0323 seconds
- **Average Similarity:** 0.5517
- **Utility Score:** 0.5194

Although the FAISS system demonstrated lower query latency, the decreased similarity scores reveal its limitations in capturing nuanced semantic relationships without a latent re-ranking component. For example:

- *"What is the process of photosynthesis and why is it important?"* — Similarity Score: 0.8708
- Remaining results — Similarity Scores: Below 0.70, with some as low as 0.38

These lower similarity values suggest that while FAISS is computationally efficient, it often compromises on retrieval quality, especially for queries requiring deeper semantic understanding.

C. Game-Theoretic Outcome

When comparing utility scores under the balanced setting of $\alpha = \beta = 1.0$, the hybrid system outperformed the FAISS-based system by a margin of 0.3679. The higher similarity

scores outweighed the slightly longer query latency, highlighting that semantic accuracy contributes more significantly to overall utility than raw retrieval speed—particularly in use cases where the quality of results is paramount.

Therefore, under the given constraints and evaluation framework, the **Custom DB (Autoencoder + HNSW)** is identified as the **dominant strategy** for vector-based semantic search from a game-theoretic perspective.

V. Conclusion

This paper demonstrates that a game-theoretic approach to latent-space compression, leveraging deep autoencoders and hybrid HNSW indexing, significantly enhances the semantic accuracy and utility of transformer-based vector search systems compared to traditional methods like FAISS. By modeling the trade-off between retrieval accuracy and storage efficiency as a zero-sum game, our proposed framework achieves near-lossless semantic retrieval in compressed spaces, with a substantial improvement in average similarity and overall utility, albeit with a modest increase in query time. These results highlight the practical value of game-theoretic optimization for scalable, high-utility information retrieval, and pave the way for more intelligent integration of compression and search strategies in future large language model pipelines.

References

[1] K. Agrawal and N. Nargund, *Deep Learning in Industry 4.0: Transforming Manufacturing Through Data-Driven Innovation*. Springer Nature Switzerland, 2024, p. 222–236. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-50583-615

[2] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020. [Online]. Available: http://dx.doi.org/10.18653/v1/2020.emnlp-main.550

[3] T. Kenter and M. de Rijke, "Short text similarity with word embeddings," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, ser. CIKM'15. ACM, oct 2015, p. 1411–1420. [Online]. Available: http://dx.doi.org/10.1145/2806416.2806475

[4] R. Xia, Y. Pan, H. Lai, C. Liu, and S. Yan, "Supervised hashing for image retrieval via image representation learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 28, no. 1, jun 2014. [Online]. Available: http://dx.doi.org/10.1609/aaai.v28i1.8952

[5] M. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," in *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, ser. MLSDA'14. ACM, dec 2014, p. 4–11. [Online]. Available: http://dx.doi.org/10.1145/2689746.2689747

[6] M. Alkhayrat, M. Aljnidi, and K. Aljoumaa, "A comparative dimensionality reduction study in telecom customer segmentation using deep learning and pca," *Journal of Big Data*, vol. 7, no. 1, feb 2020. [Online]. Available: http://dx.doi.org/10.1186/s40537-020-0286-0

[7] A. Vahdat, K. Kreis, and J. Kautz, "Score-based generative modeling in latent space," 2021. [Online]. Available: http://arxiv.org/pdf/2106.05931

[8] D. He, Z. Yang, W. Peng, R. Ma, H. Qin, and Y. Wang, "Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5708–5717.

[9] Y. Yang, H. Sun, Y. Zhang, T. Zhang, J. Gong, Y. Wei, Y.-G. Duan, M. Shu, Y. Yang, D. Wu, and D. Yu, "Dimensionality reduction by umap reinforces sample heterogeneity analysis in bulk transcriptomic data," jan 2021. [Online]. Available: http://dx.doi.org/10.1101/2021.01.12.426467

[10] T. L. Griffiths and M. Steyvers, *A probabilistic approach to semantic representation*. Routledge, apr 2019, p. 381–386. [Online]. Available: http://dx.doi.org/10.4324/9781315782379-102

[11] M.-H. In and O. Speck, "Highly accelerated psf-mapping for epi distortion correction with improved fidelity," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 25, no. 3, p. 183–192, aug 2011. [Online]. Available: http://dx.doi.org/10.1007/s10334-011-0275-6

[12] F. Trozzi, X. Wang, and P. Tao, "Umap as a dimensionality reduction tool for molecular dynamics simulations of biomacromolecules: A comparison study," *The Journal of Physical Chemistry B*, vol. 125, no. 19, p. 5022–5034, may 2021. [Online]. Available: http://dx.doi.org/10.1021/acs.jpcb.1c02081

[13] B. Hernandez, O. Stiff, D. K. Ming, C. Ho Quang, V. Nguyen Lam, T. Nguyen Minh, C. Nguyen Van Vinh, N. Nguyen Minh, H. Nguyen Quang, L. Phung Khanh, T. Dong Thi Hoai, T. Dinh The, T. Huynh Trung, B. Wills, C. P. Simmons, A. H. Holmes, S. Yacoub, and P. Georgiou, "Learning meaningful latent space representations for patient risk stratification: Model development and validation for dengue and other acute febrile illness," *Frontiers in Digital Health*, vol. 5, feb 2023. [Online]. Available: http://dx.doi.org/10.3389/fdgth.2023.1057467

[14] K. Agrawal, P. Goktas, B. Sahoo, S. Swain, and A. Bandyopadhyay, *IoT-Based Service Allocation in Edge Computing Using Game Theory*. Springer Nature Switzerland, dec 2024, p. 45–60. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-81404-44

[15] H.-Y. Shi, W.-L. Wang, N.-M. Kwok, and S.-Y. Chen, "Game theory for wireless sensor networks: A survey," *Sensors*, vol. 12, no. 7, p. 9055–9097, jul 2012. [Online]. Available: http://dx.doi.org/10.3390/s120709055

[16] L. Han, T. Morstyn, and M. McCulloch, "Incentivizing prosumer coalitions with energy management using cooperative game theory," *IEEE Transactions on Power Systems*, vol. 34, no. 1, pp. 303–313, 2019.

[17] S. Bai, P. Tang, P. H. Torr, and L. J. Latecki, "Re-ranking via metric fusion for object retrieval and person re-identification," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 740–749.

[18] C. Daskalakis and I. Panageas, "Last-iterate convergence: Zero-sum games and constrained min-max optimization." Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2019. [Online]. Available: https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2019.27

[19] J. Lin, X. Ma, S.-C. Lin, J.-H. Yang, R. Pradeep, and R. Nogueira, "Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '21. ACM, jul 2021, p. 2356–2362. [Online]. Available: http://dx.doi.org/10.1145/3404835.3463238

[20] G. Bacci, S. Lasaulce, W. Saad, and L. Sanguinetti, "Game theory for networks: A tutorial on game-theoretic tools for emerging signal processing applications," *IEEE Signal Processing Magazine*, vol. 33, no. 1, pp. 94–119, 2016.

[21] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, "Stanford alpaca: An instruction-following llama model," https://github.com/tatsu-lab/stanford$_alpaca$, 2023.