

Andrew H. Bond
Department of Computer Engineering
San José State University
San José, CA 95192, USA

April 8, 2026

Editor-in-Chief
IEEE Transactions on Artificial Intelligence

Dear Editor,

I am submitting the manuscript “PCA-Matryoshka: Enabling Effective Dimension Reduction for Non-Matryoshka Embedding Models with Applications to Vector Database Compression” for consideration in IEEE Transactions on Artificial Intelligence.

Summary. Most deployed embedding models (BGE-M3, Cohere Embed, older OpenAI models) cannot be effectively truncated for compression because they were not trained with Matryoshka representation learning. We show that a training-free PCA rotation recovers the truncation property, improving cosine similarity from 0.467 to 0.974 at 256 dimensions (109% improvement). Combined with 3-bit scalar quantization, the pipeline achieves $27\times$ compression at 0.979 cosine similarity on a 2.4-million-vector corpus spanning 37 languages.

Contributions.

1. A training-free technique (PCA-Matryoshka) that enables effective dimension truncation for any embedding model.
2. The first systematic comparison of 15 embedding compression methods on a single large-scale corpus, identifying Pareto-optimal configurations.
3. An open-source implementation deployed in production on 3.3 million vectors.

Relevance to IEEE TAI. This work bridges the gap between embedding model research and practical deployment constraints. The compression technique directly enables local AI deployment in privacy-sensitive domains (healthcare, education, legal)—a growing concern as AI regulation tightens globally.

The manuscript is fully anonymized for double-anonymous review. It has not been submitted elsewhere and is not under consideration at any other journal.

Thank you for your consideration.

Sincerely,

Andrew H. Bond, Senior Member,
IEEE
Department of Computer Engineering
San José State University
andrew.bond@sjsu.edu