Analytical Integration of Harmonic Token Projection (HTP) with FAISS for Deterministic Recommender Systems

Tcharlies Schmitz, **Débora de Faria Ferreira Gomes**
Data Science - PX.Center
Joinville, SC, Brazil
E-mail: tcharlies.schmitz@px.br, debora.gomes@px.center

Talia Correia Schulz
Postgraduate Program in Numerical Methods and Engineering – UFPR
Curitiba, PR, Brazil
E-mail: talia.correia@ufpr.br

ABSTRACT

The *Harmonic Token Projection* (HTP) [1], encodes categorical and textual attributes—such as vehicle type, cargo, supply, and contract location—into harmonic vectors derived from the integer identities of each token. Each identifier r_i is first mapped to a residue in a modular base m_i, preserving its discrete identity within the finite cyclic group $\mathbb{Z}_{m_i}$. The corresponding harmonic embedding vector $\mathbf{E_i}$ is then defined as:

$$\mathbf{E_i} = \left[\,\sin\!\left(\tfrac{2\pi r_i}{m_i}\right),\ \cos\!\left(\tfrac{2\pi r_i}{m_i}\right)\right], \tag{1}$$

where r_i represents the token's integer or Unicode index and m_i denotes the chosen modulus (typically prime or coprime across dimensions). This formulation, formalized analytically in [1], projects each token deterministically onto the unit circle, generating a phase-stable and bijective encoding. Because both sine and cosine are periodic with 2π, the mapping ensures reversibility: the original identifier can be reconstructed from the harmonic phase

$$\phi_i = \tan^{-1}\left(\frac{\sin(2\pi r_i/m_i)}{\cos(2\pi r_i/m_i)}\right) \mod m_i.$$

Thus, the representation is fully analytic, reversible, and free from floating-point drift or stochastic initialization.

These deterministic embeddings are indexed and compared through the *Facebook AI Similarity Search* (FAISS) library [3, 4]. Given a query vector $\mathbf{q}$ (active contract) and a database of source vectors $\mathbf{s}_j$ (drivers with past contracts), FAISS retrieves the top-k candidates that maximize the cosine similarity:

$$\cos(\theta) = \frac{\mathbf{q}\cdot\mathbf{s}_j}{\|\mathbf{q}\|\|\mathbf{s}_j\|}, \tag{2}$$

where $\mathbf{q}\cdot\mathbf{s}_j$ is the dot product and $\|\mathbf{q}\|$ the Euclidean norm. This analytical comparison eliminates the need for neural embeddings [5, 6, 2, 7] or learned parameters while ensuring sub-millisecond retrieval for thousands of contracts simultaneously.

The FAISS index performs a high-dimensional search, returning the 1,000 most similar historical driver embeddings. Subsequently, a vectorized Haversine function computes the geographic distance between query and source pairs:

$$d_{km} = 2R\arctan\!\left(\sqrt{a}, \sqrt{1-a}\right), \quad a = \sin^2\frac{\Delta\phi}{2} + \cos(\phi_1)\cos(\phi_2)\sin^2\frac{\Delta\lambda}{2}, \tag{3}$$

where $R = 6371.0088$ km is the Earth's mean radius [8]. A dynamic filtering stage applies adaptive radius thresholds proportional to contract complexity levels, ensuring context-aware geographic validation.

As an illustrative example, consider a new freight request requiring a "Truck" vehicle and originating from a specific geographic region. The token "Truck" is mapped to an integer identifier r, which is then projected into its harmonic representation $\mathbf{E} = [\sin(2\pi r/m), \cos(2\pi r/m)]$. Other attributes such as cargo type or contract modality are encoded in the same harmonic space, producing a compact and reversible embedding of the full operational context. The resulting query vector combines these harmonic components to form a deterministic representation of the target contract.

Using this vector, FAISS performs a cosine-based nearest-neighbor search over a database of historical embeddings, retrieving past contracts with similar operational characteristics. Geographic distance is then evaluated via the Haversine formula to ensure spatial compatibility. This two-stage process—harmonic encoding followed by geometric retrieval—demonstrates how symbolic operational data can be transformed into analytical geometric structures suitable for fast, interpretable, and reproducible recommendation workflows.

The experiments conducted with the Harmonic Token Projection (HTP) in the context of PX.Center indicate that large-scale driver–freight matching can be approached through deterministic vector geometry rather than stochastic optimization. These experiments were carried out in an industrial logistics environment characterized by high-dimensional categorical structures and heterogeneous geographic information. The results show that modular harmonic embeddings, when coupled with FAISS-based cosine similarity search [3], form a scalable and analytically tractable framework for nearest-neighbor retrieval. This deterministic pipeline provides a reversible and mathematically interpretable alternative to neural embedding models, reinforcing the feasibility of building explainable and reproducible recommendation systems in real-world operational settings.

Keywords: *Harmonic Token Projection, Deterministic embeddings, FAISS, Cosine similarity, Recommender systems, Industrial AI, Explainability.*

References

[1] T. Schmitz, Harmonic Token Projection: Analytical Encodings for Deterministic Vector Geometry, *arXiv preprint arXiv:2511.20665*, 2025.

[2] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, *NAACL-HLT*, 2019.

[3] J. Johnson, M. Douze, and H. Jégou, Billion-scale similarity search with GPUs, *IEEE Trans. Big Data*, 7(3), 535–547, 2021.

[4] H. Jégou, M. Douze, and C. Schmid, Product quantization for nearest neighbor search, *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(1), 117–128, 2011.

[5] T. Mikolov, K. Chen, G. Corrado and J. Dean, Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781*, 2013.

[6] J. Pennington, R. Socher and C. D. Manning, GloVe: Global vectors for word representation, *EMNLP*, pp. 1532–1543, 2014.

[7] N. Reimers and I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT networks, *EMNLP*, 2019.

[8] J. P. Snyder, Map projections: A working manual, *U.S. Geological Survey Professional Paper 1395*, 1987.