Harmonic Token Projection (HTP): A Vocabulary-Free, Training-Free, Deterministic, and Reversible Embedding

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

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) is introduced as a fully deterministic, reversible, and training-free framework for text representation. Unlike neural embeddings that depend on stochastic optimization and large training corpora, HTP encodes each token analytically as a harmonic trajectory derived from its Unicode integer identity. Through modular arithmetic and trigonometric projection, the method establishes a bijective mapping between discrete symbolic sequences and continuous vector space, ensuring complete reversibility and interpretability. Each token is represented as a composition of sinusoidal components corresponding to modular residues, forming a compact and mathematically transparent embedding that preserves the entire discrete structure of the input. By unifying number theory, signal processing, and geometric analysis, HTP provides a foundation for interpretable, reversible, and computationally efficient representations, bridging symbolic and continuous paradigms in artificial intelligence.

Keywords: *Harmonic Token Projection, Reversible Embedding, Deterministic Representation, Modular Arithmetic, Semantic Similarity, Explainable AI*

1 Introduction

The representation of language in continuous vector spaces has been one of the most influential developments in modern artificial intelligence. Techniques such as Word2Vec, GloVe, and transformer-based architectures (e.g., BERT and Sentence-BERT) have enabled models to capture complex semantic relations across languages and domains [2, 3, 8, 9]. However, these approaches depend heavily on large training corpora, stochastic optimization, and billions of learned parameters—resulting in systems that are computationally expensive, difficult to reproduce, and fundamentally opaque in their internal mechanics.

The inherent limitations of such data-driven embeddings motivate the search for methods that are both interpretable and deterministic. In particular, most neural embeddings are not analytically reversible: they approximate meaning through statistical co-occurrence rather than through explicit mathematical transformations. Consequently, the resulting vector spaces are emergent rather than designed, making it difficult to trace how discrete symbols map to continuous representations or to guarantee reproducibility across environments.

The *Harmonic Token Projection* (HTP) addresses these limitations by redefining text representation as a purely analytical process. Instead of relying on training or contextual prediction, HTP encodes tokens through harmonic functions derived directly from their Unicode integer

identities, creating a bijective and reversible mapping between symbolic and continuous domains [6]. Each token is projected into a continuous space using modular arithmetic and trigonometric transformation, ensuring that all information about the original symbol is preserved exactly. This geometric formulation transforms language into a field of harmonic trajectories—where meaning emerges not from probabilistic learning but from deterministic structure.

By integrating number theory, signal processing, and algebraic geometry [7], HTP bridges symbolic and sub-symbolic paradigms in a mathematically transparent way. The resulting embeddings are language-agnostic, training-free, and reversible, enabling rapid computation with negligible memory usage. Empirical evaluations on the Semantic Textual Similarity Benchmark (STS-B) and its multilingual extension demonstrate that HTP attains correlations comparable to transformer-based embeddings while requiring three orders of magnitude less computational cost [4].

The general formulation of HTP first appeared in the preprint [1]. The current work synthesizes those initial concepts into a unified methodological framework. Section 2 details the analytical methodology underlying the harmonic encoding and its reversibility; Section 3 discusses empirical results and computational efficiency; Section 4 provides a conceptual discussion situating HTP within the broader context of representation learning; and Section 5 concludes by outlining the implications of this deterministic approach for future research in interpretable and efficient artificial intelligence.

2 Methodology

The *Harmonic Token Projection* (HTP) introduces a fully deterministic and reversible process for transforming discrete text tokens into continuous numerical vectors. Unlike neural embeddings that rely on stochastic optimization or statistical co-occurrence, HTP performs an exact analytic mapping between symbolic and geometric domains. Its mathematical formulation integrates number theory, trigonometric projection, and the Chinese Remainder Theorem (CRT) to ensure bijectivity, interpretability, and computational efficiency.

Direct Process

Let a token $t = [c_1, c_2, \ldots, c_\ell]$ be a sequence of characters of length ℓ. Each character c_i is mapped to its Unicode integer code point:

$$u_i = \mathrm{ord}(c_i), \quad i = 1, 2, \ldots, \ell. \tag{1}$$

The sequence is zero-padded up to a fixed length $L_{\max}$ to ensure dimensional consistency:

$$\tilde{u} = [u_1, u_2, \ldots, u_\ell, 0, \ldots, 0], \quad \mathrm{len}(\tilde{u}) = L_{\max}. \tag{2}$$

This sequence is interpreted as a base-B integer identifier:

$$N_t = \sum_{j=1}^{L_{\max}} \tilde{u}_j \, B^{L_{\max}-j}, \quad B = 2^{16}. \tag{3}$$

The integer N_t is then decomposed into residues with respect to a set of pairwise coprime moduli $\{m_1, m_2, \ldots, m_k\}$:

$$r_i = N_t \bmod m_i, \quad i = 1, 2, \ldots, k. \tag{4}$$

Each residue defines a phase component on the unit circle and is projected harmonically:

$$E_i = [\sin(2\pi r_i/m_i), \cos(2\pi r_i/m_i)]. \tag{5}$$

The complete embedding vector is formed by concatenation of all harmonic pairs:

$$E(t) = [E_1, E_2, \ldots, E_k] \in \mathbb{R}^{2k}. \tag{6}$$

This transformation encodes the entire discrete information of t into a smooth, periodic, and reversible continuous representation.

Inverse Process

Given a harmonic embedding $E(t)$, each pair (s_i, c_i) corresponds to an angular phase:

$$\tilde{r}_i = \text{round}\left(\frac{\text{atan2}(s_i, c_i)}{2\pi}\, m_i\right) \bmod m_i. \tag{7}$$

The integer N_t is reconstructed via the Chinese Remainder Theorem:

$$N_t = \left(\sum_{i=1}^{k} \tilde{r}_i M_i y_i\right) \bmod M, \quad M = \prod_{i=1}^{k} m_i,\ M_i = \frac{M}{m_i}, \tag{8}$$

where $y_i = M_i^{-1} \bmod m_i$ is the modular multiplicative inverse of M_i. This guarantees a one-to-one correspondence between the discrete integer space and the harmonic vector space, ensuring full reversibility.

Mathematical Properties

The HTP framework satisfies five defining properties:

1. **Determinism:** identical tokens yield identical embeddings, without randomness or learned weights.
2. **Continuity:** small variations in N_t produce smooth angular changes in $E(t)$.
3. **Reversibility:** CRT-based inversion ensures exact bijective recovery within $[0, M)$.
4. **Geometric Periodicity:** each $(\sin, \cos)$ pair encodes a bounded rotation on the unit circle.
5. **Interpretability:** every coordinate has a defined analytic meaning as a harmonic of a modular residue.

Together, these properties establish a transparent, mathematically grounded bridge between symbolic sequences and continuous representations.

Harmonic Pooling

For sentence-level representation, individual token embeddings are aggregated through harmonic energy pooling — a deterministic analogue of TF–IDF weighting [3]. Each token t_i is weighted by its inverse token frequency:

$$w(t_i) = \frac{1}{\log(1 + f(t_i))}, \tag{9}$$

where $f(t_i)$ denotes its frequency in a reference corpus. The sentence embedding is computed as a normalized weighted mean:

$$v(S) = \frac{\sum_{i=1}^{n} w(t_i) E(t_i)}{\sum_{i=1}^{n} w(t_i)}, \quad v'(S) = \frac{v(S)}{\|v(S)\|_2}. \tag{10}$$

This operation maintains the harmonic coherence of the representation and supports efficient cosine-based similarity:

$$\text{sim}(x, y) = \frac{\langle x, y \rangle}{\|x\|_2 \, \|y\|_2}. \tag{11}$$

The pooling mechanism acts as a geometric smoothing filter, attenuating frequent words while emphasizing semantically informative ones. Its analytic nature guarantees reproducibility and enables efficient implementation with linear time complexity $O(|S| \cdot k)$.

The *Harmonic Token Projection* (HTP) introduces a novel paradigm for analytical text representation, showing that deterministic geometry can approximate semantic structures without relying on statistical learning. Its formulation merges symbolic computation, number theory, and trigonometric analysis, achieving a synthesis between discrete logic and continuous geometry while maintaining complete reversibility, interpretability, and minimal computational cost.

The HTP model is founded on analytical determinism rather than empirical optimization. Each stage—from Unicode mapping to harmonic projection—was designed to preserve bijectivity and mathematical transparency. By encoding tokens as harmonic trajectories on the unit circle, the method treats language as a structured geometric field instead of a stochastic distribution of co-occurrences. This ensures a fully interpretable embedding space in which each coordinate corresponds to a modular residue that can be inverted exactly using the Chinese Remainder Theorem. The use of Unicode as a semantic coordinate system provides universality, assigning each symbol a unique integer identity. This eliminates the ambiguity of corpus-dependent vocabularies and allows language-agnostic and multilingual embeddings, as demonstrated by Schmitz [6].

Operating purely analytically, HTP requires no stochastic updates or gradient-based optimization. Its computational complexity scales linearly with text length and the number of harmonic components, enabling efficient computation on standard CPUs. The absence of randomness eliminates rounding instability and cumulative numerical drift, ensuring that identical input always produces identical embeddings across environments. This determinism redefines reproducibility in representation learning and guarantees that every encoded element can be precisely reconstructed—an uncommon property in modern embedding systems.

HTP also reveals that the internal structure of symbolic representation can emerge directly from geometry. Each vector dimension corresponds to a harmonic component, offering a level of interpretability rarely seen in stochastic models. In contrast to neural embeddings that learn associations implicitly, HTP derives meaning through analytical symmetry and periodicity, aligning with the principles of explainable artificial intelligence (XAI). This supports the broader insight that semantic similarity may arise not only from contextual frequency but also from intrinsic topological relationships among symbols [7].

Although deterministic, HTP can serve as the foundation for hybrid architectures. Neural embedding layers initialized with harmonic projections can benefit from structured and reversible coordinate systems that accelerate convergence while preserving interpretability. This combination of analytical precision and adaptive contextualization may yield models that are both efficient and transparent. Beyond natural language processing, HTP can generalize to other discrete domains such as genomic sequence encoding, reversible database indexing, cryptographic hashing, and symbolic compression—domains in which traceability and reversibility are essential.

Despite its coherence, HTP lacks contextual disambiguation: words with multiple meanings receive identical encodings, as the model captures form rather than use. Moreover, linear pooling may weaken compositional semantics in longer sequences. Future work should explore adaptive frequency weighting, phase-aware pooling, and multi-scale harmonic bases, as well as hybrid Fourier–neural architectures that preserve deterministic cores while learning contextual refinements. Such developments could bridge analytical transparency and the adaptive capacity of deep learning, advancing toward interpretable and efficient representation systems.

3 Conclusion

The *Harmonic Token Projection* (HTP) establishes a mathematically rigorous alternative to conventional data-driven embeddings, demonstrating that deterministic geometry can approximate semantic similarity without reliance on stochastic optimization or large training corpora. By combining modular arithmetic, trigonometric projection, and the Chinese Remainder Theorem into a unified analytic framework, HTP achieves complete reversibility, high interpretability,

and computational efficiency rarely observed in modern representation models.

The proposed method reframes text encoding as a bijective transformation between symbolic and continuous domains. Each token is treated as a harmonic oscillator whose phase space is determined directly by its Unicode integer identity, producing a continuous and reversible mapping that preserves all discrete information. This approach eliminates the dependence on vocabulary tables, statistical co-occurrence, and parameter learning—yielding embeddings that are fully reproducible and language-agnostic. Beyond its immediate performance, the conceptual contribution of HTP lies in restoring analytical transparency to the process of representation learning. It demonstrates that reversibility, interpretability, and efficiency need not be mutually exclusive—a principle that redefines the theoretical boundary between symbolic computation and continuous representation. As a result, HTP provides both a practical encoding scheme and a mathematical foundation for the design of future deterministic architectures in artificial intelligence.

Future research directions include exploring multi-scale harmonic embeddings, adaptive frequency modulation, and hybrid systems where deterministic initialization serves as a scaffold for contextual neural fine-tuning. Such extensions could unify analytic precision with the adaptive expressiveness of deep learning, paving the way for a new generation of models that are not only efficient but also inherently interpretable and verifiable.

In essence, the *Harmonic Token Projection* represents a paradigm shift: it transforms embedding from an empirical art into an analytical science—where geometry, arithmetic, and language converge into a single reversible structure capable of explaining and reproducing meaning with mathematical clarity.

Acknowledgments

This research was supported by **PX.Center** — a Brazilian logistics platform focused on freight brokerage and transportation optimization (`https://px.center`). The PX.Center provided computational infrastructure, datasets, and a research environment that enabled the development and validation of this study.

4 References

References

[1] T. Schmitz *Harmonic Token Projection (HTP): A Vocabulary-Free, Training-Free, Deterministic, and Reversible Embedding*, arXiv preprint arXiv:2511.20665, 2025.

[2] C. Guo and F. Berkhahn, Entity embeddings of categorical variables, *arXiv preprint arXiv:1604.06737*, 2016.

[3] K. D. Mwamba and I. Joe, A deep-learned embedding technique for categorical features encoding, *IEEE Access*, vol. 9, pp. 114 587–114 598, 2021.

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

[5] K. Rosen, *Elementary Number Theory and Its Applications*, 5th Edition, Addison Wesley, 2004.

[6] T. Schmitz, Modular Linear Tokenization (MLT), *arXiv preprint arXiv:2510.25952*, 2025.

[7] R. Tolimieri, M. An, and C. Lu, *Mathematics of Multidimensional Fourier Transform Algorithms*, Springer, 1998.

[8] M. Wang, S. Chen, and D. Wüthrich, Machine learning with high-cardinality categorical features in actuarial applications, *ASTIN Bulletin: The Journal of the IAA*, 54(3), pp. 689–715, 2024.

[9] A. Wilson, J. Lee, and R. Huang, Efficient representations for high-cardinality categorical variables, *arXiv preprint arXiv:2501.05646*, 2025.