Modular Linear Tokenization: A Deterministic and Reversible Framework for Categorical Data Encoding over Finite Fields.

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

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

Abstract: The Modular Linear Tokenization (MLT) is presented as a deterministic and reversible framework for categorical data representation based on modular arithmetic and linear algebra over finite fields. The methodology defines a bijective mapping between discrete identifiers and compact numerical vectors, ensuring exact reversibility and eliminating collisions typical of stochastic embeddings. By combining prime-base decomposition and modular linear transformations, MLT achieves constant computational complexity with respect to vocabulary size, enabling scalability to millions of identifiers. The approach provides full mathematical traceability and numerical stability, making it suitable for machine learning pipelines that demand reproducibility, interpretability, and integrity in categorical encodings. Geometric and computational analyses demonstrate that MLT bridges symbolic and continuous representations, establishing a foundation for future deterministic architectures in artificial intelligence.

Keywords: *Modular Linear Tokenization, Deterministic Embeddings, Modular Arithmetic, Finite Fields, Reversible Encoding, Machine Learning*

1 Introduction

The numerical representation of categorical data remains a central challenge in applied mathematics and machine learning. Traditional approaches—such as one-hot encoding, hashing, or stochastic embeddings—[1, 2, 4, 5] often involve trade-offs among dimensional efficiency, reversibility, and interpretability. In high-cardinality domains, where the number of unique identifiers can reach millions, these limitations become particularly restrictive, leading to excessive storage demands, potential information loss, and non-deterministic behavior across models.

The *Modular Linear Tokenization* (MLT) framework addresses this problem by redefining categorical encoding as a fully algebraic process grounded in *modular arithmetic* and *linear algebra over finite fields*. By treating categorical identifiers as mathematical objects rather than statistical tokens, MLT provides a deterministic and reversible alternative to stochastic embedding methods [3]. The framework establishes a bijective mapping between discrete indices and compact numerical vectors, ensuring that every encoded vector can be uniquely and exactly decoded back to its original identifier—thereby eliminating collisions and preserving full traceability.

Conceptually, MLT bridges symbolic representation and continuous computation, offering a framework that combines algebraic rigor with computational efficiency. This makes it particularly suitable for large-scale machine learning pipelines that require reproducibility, interpretability, and consistency in categorical encoding.

This paper presents the theoretical formulation, computational design, and interpretative implications of the MLT framework. The *Methodology* section formalizes the algebraic principles underlying the approach, followed by a *Discussion* that explores its conceptual scope, scalability, and integration potential within modern data-driven systems. Finally, the *Conclusion* synthesizes the main contributions and outlines directions for future research on deterministic and interpretable data representations.

2 Methodology

The *Modular Linear Tokenization* (MLT) was conceived as a deterministic and reversible methodology for representing high-cardinality categorical identifiers as compact numerical vectors. Its formulation integrates principles of modular arithmetic, invertible linear transformations, and prime-base decomposition, ensuring a bijective correspondence between the original discrete space and its projected vector representation. This section formalizes the theoretical foundations, algorithmic structure, and computational properties of the proposed method.

The central problem addressed by MLT concerns the efficient and lossless representation of categorical identifiers. Given a finite set of identifiers $I = \{0, 1, \ldots, V-1\}$, where V denotes the vocabulary cardinality, the objective is to define an encoding function $f : I \to \mathbb{Z}_p^n$ such that:

$$f(i) = (M \cdot v_i) \bmod p, \tag{1}$$

where $M \in \mathbb{Z}_p^{n \times n}$, $\det(M) \neq 0 \pmod p$ and v_i denotes the v_i i-th element represented in base p.

To ensure reversibility, the inverse function f^{-1} must exist and be directly computable:

$$f^{-1}(t) = v_i = (M^{-1} \cdot t) \bmod p. \tag{2}$$

The composition property,

$$f^{-1}(f(i)) = i, \tag{3}$$

guarantees exact reconstruction, eliminating both ambiguity and collisions. The modular representation also prevents the linear dimensional growth typical of traditional encodings, while providing explicit control over computational complexity through the parameters p and n.

The selection of p as a prime number is fundamental, ensuring that $\mathbb{Z}_p$ forms a finite field where every nonzero element possesses a multiplicative inverse. This algebraic property underpins the reversibility and numerical stability of the method.

Base-p Decomposition

The first stage of the MLT process converts the integer identifier i into a vector representation v expressed in base p:

$$v = [d_0, d_1, \ldots, d_{n-1}], \tag{4}$$

where each digit $d_k \in [0, p-1]$, and

$$i = \sum_{k=0}^{n-1} d_k \, p^k. \tag{5}$$

To ensure complete representability of all identifiers, the number of possible combinations must exceed the vocabulary size:

$$p^n > V. \tag{6}$$

For example, if $V = 10^6$ and $p = 101$, then $n \geq 3$, since $101^3 = 1,030,301 > 10^6$. This base-p decomposition establishes a modular positional structure in which each digit acts as an independent coordinate within $\mathbb{Z}_p$, resulting in a structured discrete code suitable for linear operations in finite space.

Modular Linear Transformation

The second stage applies an invertible linear transformation over $\mathbb{Z}_p^n$, defined by a matrix M satisfying $\det(M) \neq 0 \pmod{p}$. The encoding is then given by:

$$t = (M \cdot v) \bmod p. \tag{7}$$

This transformation behaves analogously to a linear cryptographic mixing process, distributing the information across all vector dimensions. Consequently, even adjacent identifiers (e.g., 1000 and 1001) are mapped to highly distinct vectors, eliminating undesired numerical correlations.

In practical implementations, M can be generated pseudo-randomly under the invertibility constraint:

$$M = \text{randint}(1, p-1, (n,n)) \text{ with } \det(M) \not\equiv 0 \pmod{p}. \tag{8}$$

The inverse matrix M^{-1} is computed once and stored for decoding, using standard modular inversion algorithms (e.g., Gauss–Jordan elimination mod p). This process is entirely deterministic and parameter-free, ensuring reproducibility and mathematical stability across environments.

Decoding and Reversibility

Decoding follows the inverse procedure:

$$v = (M^{-1} \cdot t) \bmod p, \tag{9}$$

and the reconstruction of the original identifier is obtained by:

$$i = \sum_{k=0}^{n-1} v_k \, p^k. \tag{10}$$

This guarantees a one-to-one correspondence between identifiers and their encoded representations, yielding perfect reversibility with no information loss. This property distinguishes MLT from conventional techniques such as hashing, label encoding, and supervised embeddings [1, 2], which are inherently non-invertible. Moreover, MLT is collision-free: two distinct identifiers can never produce the same output vector. This follows from the theorem:

$$M \text{ invertible in } \mathbb{Z}_p \;\Rightarrow\; f(i_1) = f(i_2) \;\Rightarrow\; i_1 = i_2. \tag{11}$$

The guarantee of injectivity and reversibility makes MLT particularly valuable for applications requiring strict traceability and auditability, such as large-scale recommendation systems, logistics optimization, and healthcare data management [4, 5].

3 Numerical Example

The experimental evaluation of the Modular Linear Tokenization (MLT) framework was conducted using a simulated vocabulary of one million categorical identifiers, a scale representative of high-cardinality industrial datasets. A prime modulus $p = 1009$ and dimension $n = 2$ were selected to satisfy the constraint $p^n > 10^6$, ensuring complete representability of the identifier space. An invertible matrix $M \in \mathbb{Z}_{1009}^{2\times 2}$ and its exact modular inverse M^{-1} were generated, with validation performed through the identity $(MM^{-1}) \bmod p = I_2$, confirming full algebraic consistency of the encoding transformation. Five high-value identifiers sampled from the simulated vocabulary were encoded deterministically into two-dimensional vectors in $\mathbb{Z}_{1009}^2$, demonstrating the compactness and structural regularity afforded by the MLT formulation.

To illustrate the numerical formulation step by step, consider the identifier $i = 373\,315$. With $p = 1009$ and $n = 2$, i is written in base p as

$$i = d_0 + d_1 p, \quad \text{with} \quad d_0 = i \bmod p = 994, \;\; d_1 = \left\lfloor \frac{i}{p} \right\rfloor = 369,$$

so that the base-p vector is

$$\mathbf{v} = \begin{bmatrix} d_0 \\ d_1 \end{bmatrix} = \begin{bmatrix} 994 \\ 369 \end{bmatrix} \in \mathbb{Z}_{1009}^2.$$

The encoding step applies the modular linear transformation

$$\mathbf{t} = (M\mathbf{v}) \bmod p, \quad M = \begin{bmatrix} 774 & 125 \\ 369 & 348 \end{bmatrix},$$

yielding

$$\mathbf{t} = \begin{bmatrix} 774 & 125 \\ 369 & 348 \end{bmatrix} \begin{bmatrix} 994 \\ 369 \end{bmatrix} \bmod 1009 = \begin{bmatrix} 209 \\ 788 \end{bmatrix}.$$

Decoding uses the exact modular inverse $M^{-1} \in \mathbb{Z}_{1009}^{2\times 2}$,

$$M^{-1} = \begin{bmatrix} 188 & 292 \\ 79 & 105 \end{bmatrix}, \quad MM^{-1} \bmod 1009 = I_2,$$

and reconstructs the base-p vector via

$$\mathbf{v}' = (M^{-1}\mathbf{t}) \bmod p = \begin{bmatrix} 188 & 292 \\ 79 & 105 \end{bmatrix} \begin{bmatrix} 209 \\ 788 \end{bmatrix} \bmod 1009 = \begin{bmatrix} 994 \\ 369 \end{bmatrix},$$

which coincides exactly with $\mathbf{v}$. Finally, the original identifier is recovered by the inverse base-p mapping

$$i' = d_0' + d_1' p = 994 + 369 \cdot 1009 = 373\,315,$$

verifying that $i' = i$ and thus confirming the bijective and reversible nature of the MLT encoding for this concrete numerical instance.

This numerical walkthrough demonstrates the full determinism and transparency of the MLT pipeline: every stage—from base-p decomposition to modular linear transformation and exact decoding—can be reproduced algebraically without ambiguity or dependence on floating-point operations. The example highlights the core strengths of the method: strict bijectivity, collision-free encoding, and mathematically guaranteed reversibility, even for identifiers on the order of hundreds of thousands. By exposing the internal mechanics of the transformation in explicit numerical form, the example reinforces the operational clarity of the MLT framework and provides a concrete template for practitioners seeking to implement deterministic and interpretable categorical encodings in large-scale systems.

4 Discussion

The *Modular Linear Tokenization* (MLT) framework introduces a novel algebraic paradigm for categorical encoding, combining deterministic reversibility with strict control over representational dimensionality. The results outlined in the methodological formulation underscore its advantages over traditional stochastic or learned embedding schemes, yet several conceptual implications and practical trade-offs merit closer examination.

From a theoretical perspective, MLT bridges two domains that have historically evolved in isolation: symbolic encoding and continuous representation learning [1, 2]. By operating entirely within the finite field $\mathbb{Z}_p^n$, the method preserves the discrete nature of categorical identifiers while enabling linear operations that integrate seamlessly with modern machine learning models. This

duality confers a distinct interpretability advantage, as every transformation step is analytically invertible and auditable. In contrast, conventional embedding methods—such as Word2Vec or learned dense encoders—sacrifice exact traceability in pursuit of optimization flexibility.

The deterministic nature of MLT also brings unique benefits to reproducible research and production environments. Because each transformation is fully determined by (p, n, M) and a fixed random seed, the encoding process can be exactly replicated across systems and over time. This property is particularly valuable in large-scale data pipelines, where the consistency of categorical representations is critical for model versioning, explainability, and regulatory compliance.

However, the same algebraic rigidity that ensures reversibility also introduces practical limitations. Unlike adaptive embeddings, MLT does not learn task-specific representations; its expressive power depends entirely on the modular parameters chosen. Consequently, when applied in purely predictive contexts, MLT may require additional trainable layers to capture semantic relationships between categories. In this regard, MLT should be viewed not as a competitor to deep embeddings, but rather as a deterministic substrate upon which higher-level, data-driven representations can be learned.

From a computational standpoint, the method exhibits remarkable scalability. Its linear or near-constant complexity with respect to vocabulary size makes it deployable in industrial systems handling millions of identifiers. The simplicity of modular arithmetic ensures both numerical stability and hardware portability, allowing efficient implementations on CPUs and embedded systems alike. Furthermore, the absence of floating-point operations eliminates rounding errors—an issue frequently encountered in high-dimensional vector computations.

Looking forward, hybrid architectures that combine the deterministic backbone of MLT with learnable transformations offer a promising avenue for research. For instance, initializing a neural embedding layer with MLT-derived vectors could merge algebraic consistency with adaptive learning dynamics. Another direction involves extending MLT toward harmonic or Fourier-like formulations, wherein modular frequencies encode hierarchical or periodic categorical patterns while maintaining full invertibility.

In summary, MLT redefines the boundary between symbolic determinism and continuous learning. Its main contribution extends beyond encoding efficiency, restoring mathematical transparency to a process that, in contemporary deep learning, is often opaque. Although further empirical validation across domains is warranted, the theoretical coherence and computational elegance of MLT position it as a robust foundation for the next generation of deterministic and interpretable machine learning architectures.

5 Conclusion

The *Modular Linear Tokenization* (MLT) formalism establishes a rigorous algebraic foundation for categorical representation in computational systems [3]. By unifying modular arithmetic and linear algebra within a deterministic and reversible framework, MLT achieves what conventional encoding schemes typically cannot: exact bijectivity, numerical stability, and full auditability across large-scale data domains. Its capacity to represent millions of categorical identifiers as compact, invertible vectors provides a mathematically transparent alternative to stochastic embedding methods.

The methodology demonstrates that reversibility and efficiency are not mutually exclusive. Through the modular transformation $t = (M \cdot v) \bmod p$, MLT guarantees that each category is uniquely encoded within a finite field, while the invertible matrix M distributes information uniformly across dimensions. This structure ensures complete information preservation and deterministic reconstruction—properties that have profound implications for explainability, data integrity, and reproducibility in artificial intelligence systems.

Beyond its mathematical elegance, MLT decouples categorical encoding from statistical training, enabling fully reproducible and interpretable transformations independent of model param-

eters. This independence makes it particularly suited to industrial and scientific pipelines that demand traceable, version-controlled, and regulation-compliant feature representations—such as logistics optimization, healthcare analytics, and large-scale recommender systems.

In summary, the MLT framework represents a robust contribution to the field of mathematical tokenization, combining conceptual simplicity with computational precision. It reintroduces algebraic rigor into a domain increasingly governed by probabilistic heuristics, providing both a theoretical foundation and a practical instrument for advancing reversible, interpretable, and high-performance data representations.

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.

6 References

References

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

[2] 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.

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

[4] 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.

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