Application of Modular Linear Tokenization (MLT) in Recommender Systems

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

This work presents the application of the *Modular Linear Tokenization* (MLT) methodology in a real supervised learning scenario, validating its use in large-scale recommender systems. MLT is a deterministic and reversible technique for categorical variable encoding, based on modular arithmetic and linear transformations over finite fields [3]. Its main advantage lies in the ability to represent millions of identifiers as compact vectors, without the need for training or additional parameters, while maintaining complete reproducibility.

Traditional methods such as *One-Hot Encoding* and the *Hashing Trick* [1, 5] are widely used for categorical representation, but they either lead to extremely high dimensionality or suffer from non-reversibility and hash collisions. More recent approaches based on deep-learned embeddings [2, 4] provide efficient representations but require extensive training and lack interpretability. The MLT approach seeks to overcome these limitations by providing a mathematically grounded, fully deterministic alternative.

The methodology was applied to the *MovieLens 20M* dataset, which contains approximately 20 million user–movie interactions. The goal was to predict ratings higher than four stars, comparing MLT with traditional encoding techniques and with supervised embeddings. The evaluated metrics included vector dimensionality, average training time per epoch, inference time, and accuracy.

Method	**Dim.**	**Reversible**	**Accuracy (%)**	**Time/epoch (s)**
One-Hot	164 320	Yes	74,18	5029,9
Hashing	512	No	57,87	144,3
MLT	**14**	**Yes**	**63,31**	**65,0**
MLT+Autoenc.	16	Yes	62,53	2,1
Embeddings	32	No	74,40	410,6

Table 1: Comparative results for categorical encoding on the *MovieLens 20M* dataset.

The results show that MLT reduces dimensionality by up to 99.9% while maintaining competitive accuracy and eliminating the need for millions of trainable parameters. Its deterministic execution lowers computational training cost by up to 90%, while reversibility ensures full traceability of original identifiers—an essential advantage in industrial applications that require data integrity and auditability.

Therefore, MLT demonstrates practical applicability as a mathematically grounded and computationally efficient alternative to traditional embedding techniques [1, 2, 4]. The approach combines precision, interpretability, and scalability, making it suitable for recommender systems, credit modeling, logistics, and other applications involving large volumes of categorical variables. Future work includes harmonic projections and neural compression layers to enhance the expressiveness of encoded vectors.

Keywords: *Categorical encoding, Deterministic representations, Modular arithmetic, Recommender systems, Scalability.*

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*, 9, 114587–114598, 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*, 54(3), 689–715, 2024.

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