Regress, Don't Guess – A Regression-like Loss on Number Tokens for Language Models

Zausinger, Pennig, Chlodny, Limbach, Ketteler, Prein, Singh, Danziger, Born

IBM Research · LMU Ludwig-Maximilians-Universität München · TUM · TUM.ai

Problem Statement

- Language models (LMs) **struggle with handling numerical data effectively**, despite numbers being ubiquitous in natural texts, especially in scientific domains like chemistry and biology. The core challenges in numerically processing text include:

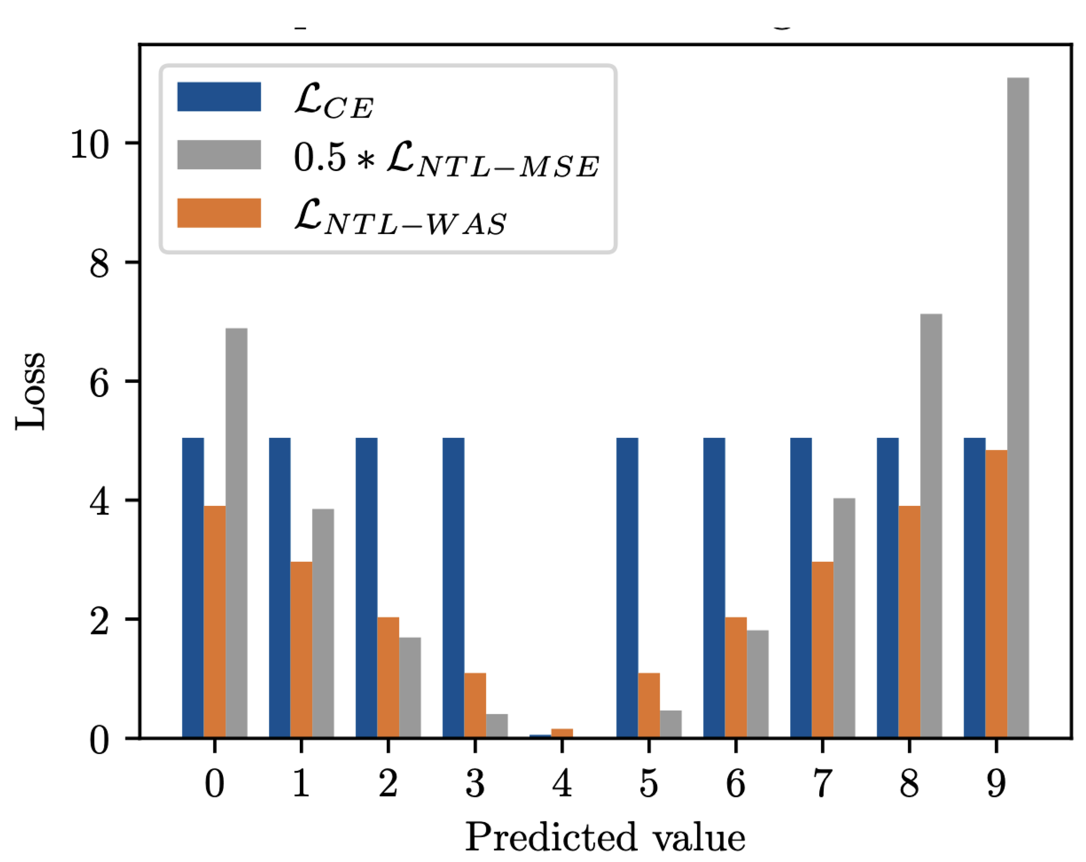
1) **Tokenization**: Standard subword tokenization breaks numbers into arbitrary tokens, disrupting their numerical structure.
2) **Embedding**: Models must recreate numerical structure from scratch, as numerical tokens are learned like any other tokens.
3) **Training Objective**: The standard cross-entropy loss treats numbers nominally, failing to capture numerical proximity. For instance, predicting 3 instead of 2 doesn't meaningfully differentiate the error.

Prior Work

Language models have relied on various strategies to address numerical limitations:

- **Cross-Entropy Loss**: The default objective for language models fails to capture numerical relationships. It treats predictions like 3 and 9 as equally incorrect when the target is 2, neglecting proximity (see Figure).
- **xVal Encoding**: This method encodes numbers as a single token with a regression head. While promising, it struggles with large numerical ranges due to its reliance on scaled embeddings and normalization layers [1].
- **Regression Transformer**: Uses digit-level tokenization with positional embeddings, maintaining numerical structure but not addressing the loss limitation [2].
- **Verifiers and calculators**: Post-hoc solutions that increase computational overhead [3].

However, these approaches fail to fundamentally improve number handling at the loss-function level.



The graph shows how CE, NTL-MSE, and NTL-WAS handle errors for predictions near the true label. CE assigns the same loss regardless of proximity, while NTL-WAS and NTL-MSE penalize based on distance, with NTL-WAS showing superior behavior.

Number Token Loss

To address the limitations of the standard Cross-Entropy (CE) loss in handling numerical data, we introduce the **Number Token Loss (NTL)**. This novel loss leverages numerical proximity and integrates seamlessly with existing language models.

We propose two variants:

1. NTL-MSE (Mean Squared Error):

- Given a model $f(\cdot)$, input tokens $x_{\leq i}$ (with $i \leq N$), the numerical value y_i' of ground truth token y_i and a vocab V consisting of tokens (with indices $j, \ldots, k$ representing the number token), we compute NTL-MSE:

$$\boxed{L_{NTL-MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i' - f(x_{\leq i})_{j:k} \circ V_{j:k})^2}$$

- The loss minimizes the squared error between the weighted sum of predicted probabilities and the ground truth value.

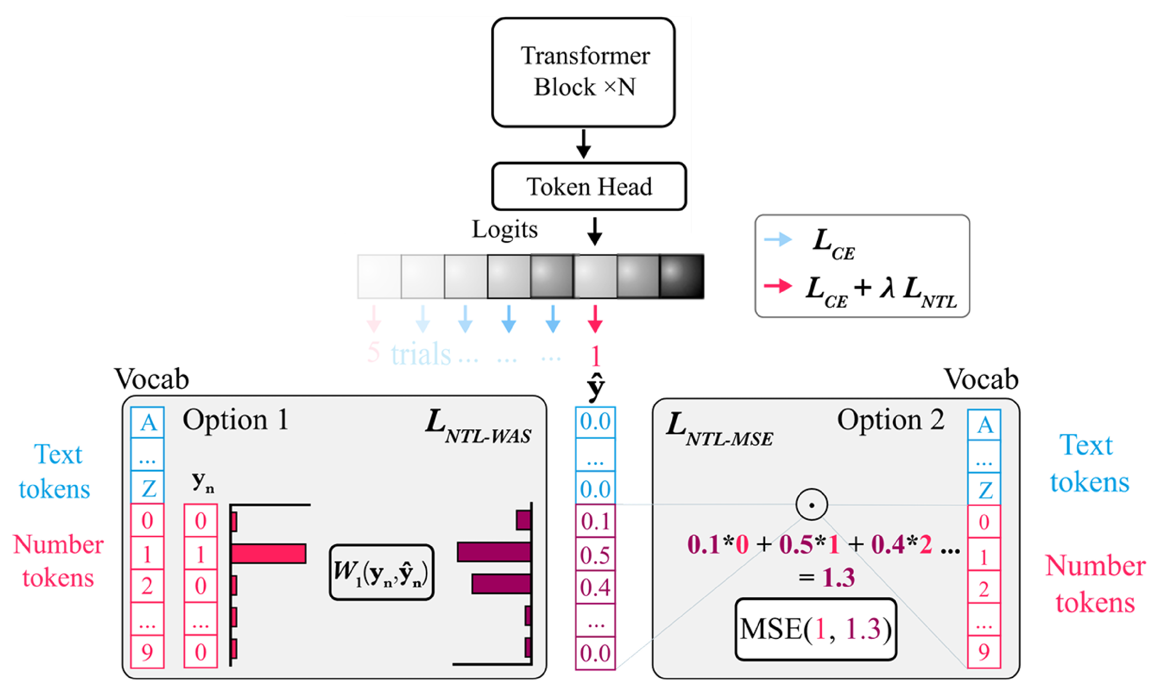
2. NTL-WAS (Wasserstein Distance):

$$\boxed{L_{NTL-WAS} = \frac{1}{N}\sum_{i=1}^{N} W_1(y_i, f(x_{\leq i})_{j:k})}$$

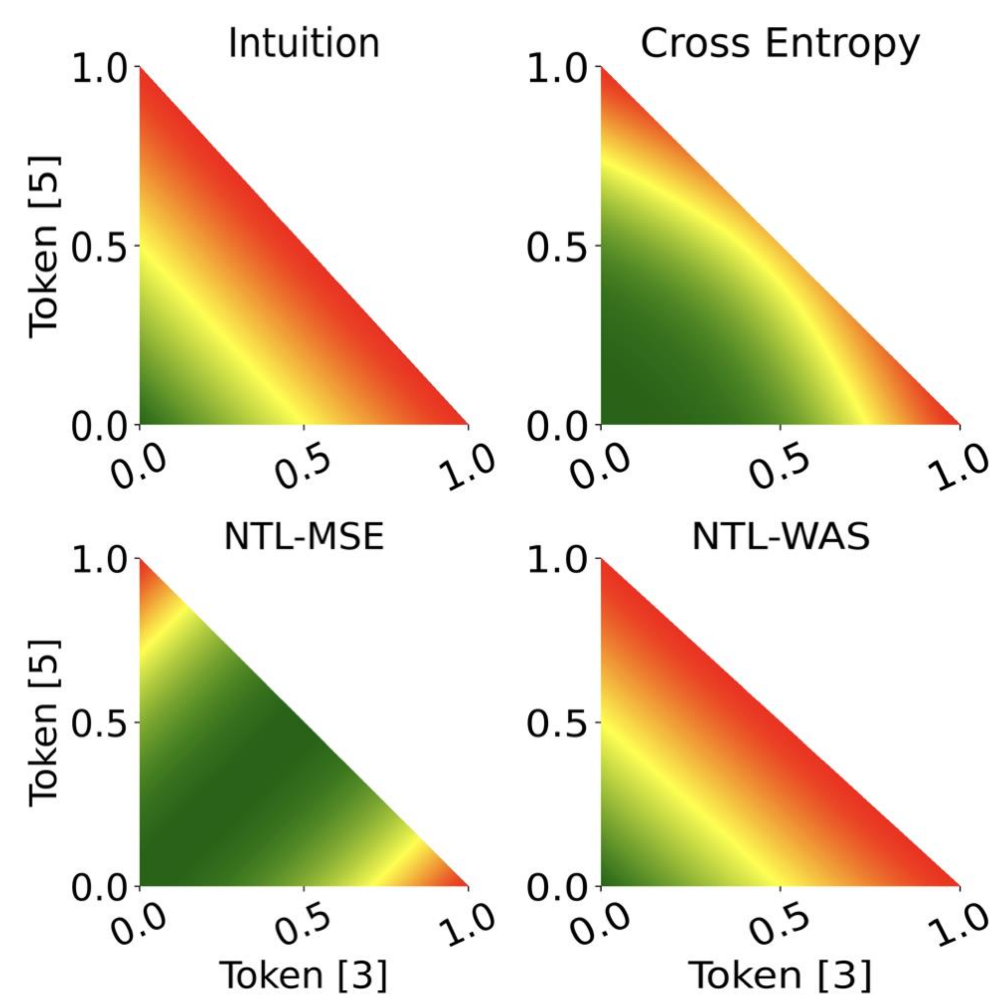
- The loss computes the Wasserstein-1 distance between the predicted and true distributions, ensuring proximity-based alignment.

Both losses are applied only to numerical tokens and integrate seamlessly with existing cross-entropy objectives, enhancing numerical reasoning without architectural changes.

$$L = L_{CE} + \lambda L_{NTL}$$



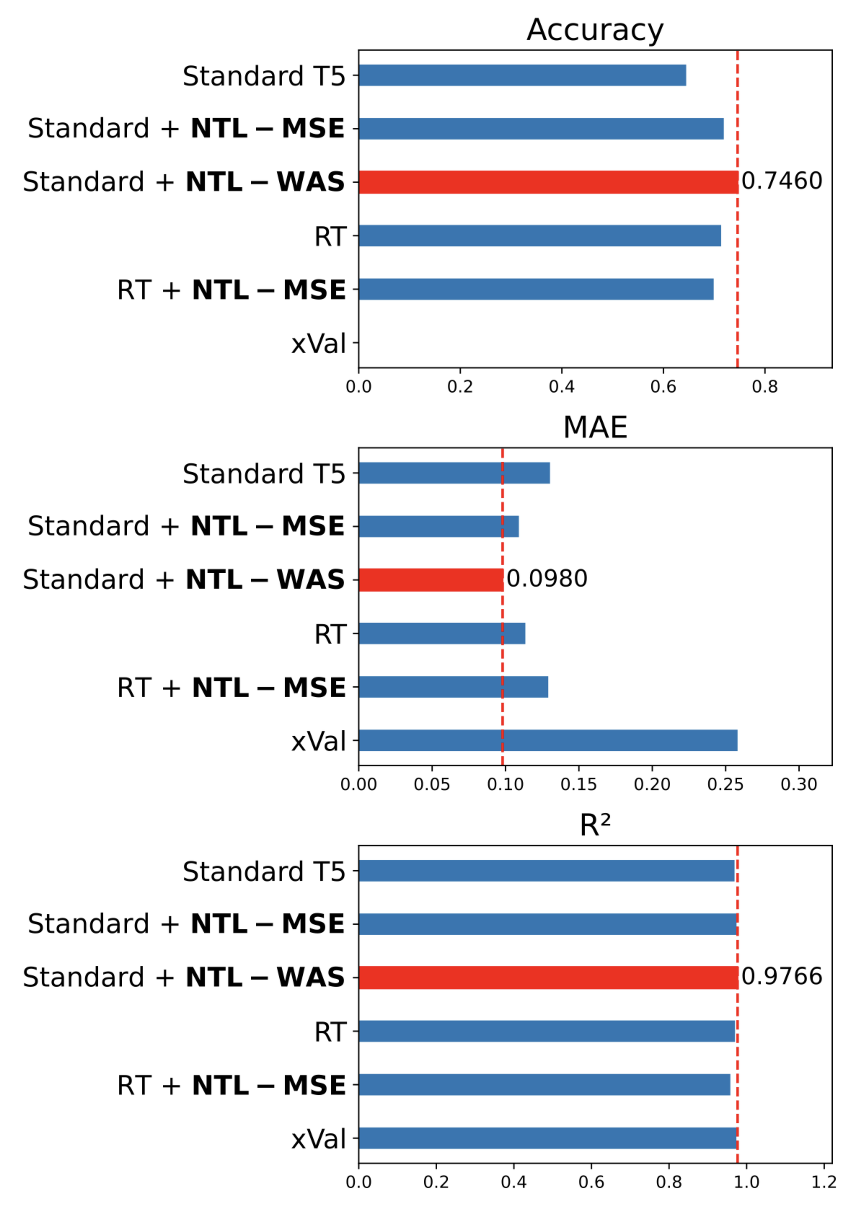
This figure illustrates the architecture for applying NTL. The transformer predicts token probabilities, which are used to compute regression-based losses.



A heatmap comparing NTL-MSE and NTL-WAS. NTL-WAS better aligns predictions with proximity-based expectations, avoiding non-injective errors.

Experiments

- **Datasets**: DeepMind Mathematics Dataset with 25M+ samples.
- **Key Results**:
  - NTL-WAS improves accuracy by about 10% in interpolation tasks over T5.
  - Significant gains in MAE and R^2, demonstrating better numerical reasoning.
  - Integration is seamless, with minimal computational overhead.



Evaluation on interpolated test data

Model	Acc.	MAE	R^2
Standard T5	.3686	0.7847	.9127
Standard + **NTL-MSE**	.4278	0.7789	.9091
Standard + **NTL-WAS**	**.4324**	**0.7438**	**.9132**
RT	.4042	0.9868	.7377
RT + **NTL-MSE**	.4282	1.0988	.6473
xVal	.0000	0.8259	.8186

Evaluation on extrapolated test data

Conclusion

NTL-WAS and NTL-MSE enable LMs to process numerical data effectively by leveraging numerical proximity. Minor loss function modifications result in significant performance improvements, particularly for numerically-rich tasks.

Future research will explore scaling NTL to large language models and extending its applicability to complex datasets involving hybrid textual-numerical representations.

Sources

[1] Siavash Golkar, Mariel Pettee, Michael Eickenberg, Alberto Bietti, Miles Cranmer, GeraudKrawezik, Francois Lanusse, Michael McCabe, Ruben Ohana, Liam Parker, et al. xval: Acontinuous number encoding for large language models. arXiv preprint arXiv:2310.02989,2023

[2] Jannis Born and Matteo Manica. Regression transformer enables concurrent sequence regressionand generation for molecular language modelling. Nature Machine Intelligence, 5(4):432–444,2023.

[3] Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen.Making language models better reasoners with step-aware verifier. In Proceedings of the 61stAnnual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),pages 5315–5333, 2023.