# Adapting Language Models via Token Translation

Zhili Feng, Tanya Marwah, Nicolò Fusi,

David Alvarez-Melis, Lester Mackey

*Carnegie Mellon University, Microsoft Research*

## Issues with existing tokenization

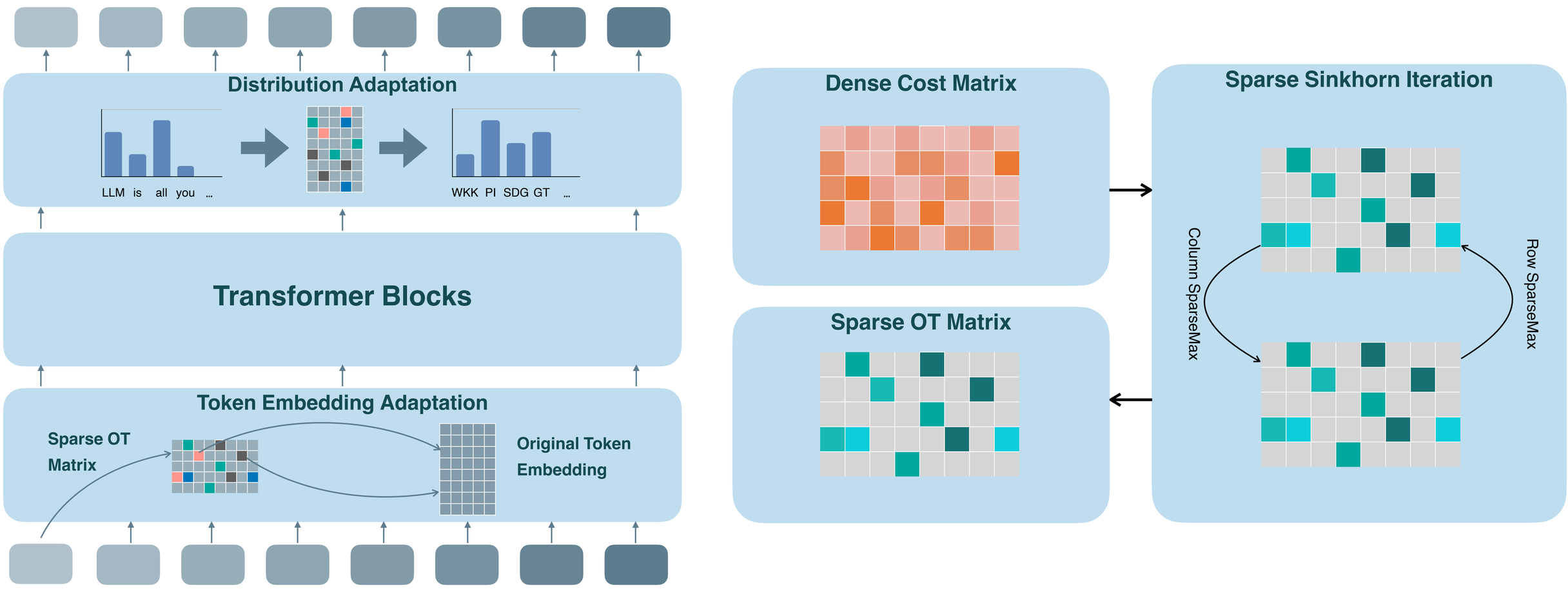
Tokenizers are trained separately from LLMs:

- Bad compression on OOD data
- Semantic misalignment: "A" in English text vs "A" in protein sequence

## This work

First retrain tokenization on OOD data, then translate tokens via sparse Sinkhorn iterations:

- Better performance on OOD data
- Robust to model upscaling
- Interpretable




**Left**: On UniRef 50, we consistently outperform baseline methods. **Right**: S2T2-k means our method is regularized with entropy penalty of strength k. The translation matrix is trained on OLMo-1B and used to initialized a OLMo-7B, on which the loss is evaluated.