



NanoTabPFN in ~10 minutes

Salih Bora Öztürk ¹ Alexander Pfefferle ^{2,1} Frank Hutter ^{3,2,1}
¹University of Freiburg ²ELLIS Institute Tübingen ³Prior Labs

TL;DR

modded-nanoTabPFN: This repository hosts the nanoTabPFN speedrun, in which we search for the fastest way to train a tabular foundation model that beats Random Forest on TabArena datasets.

Introduction

Motivation

- TabPFN have shown that pretraining on synthetic datasets can lead to strong performance... [1] [2]
- However, training these models takes long, and no one has time to wait...

Background

- GPT2 → nanoGPT → modded-nanoGPT (45 mins to 1.54 mins)
- TabPFNV2 → nanoTabPFN → modded-nanoTabPFN

Goal

- Pretrain a neural network to beat Random Forest (~0.8068 validation average ROC AUC) on subsampled TabArena datasets using 1 NVIDIA L40S.

This repo now contains a training algorithm which attains the target performance in:

- 9.26 minutes** on 1xL40S (baseline needed 74.32)
- 13184 synthetic datasets** (baseline needed 80576)

Rules

New records must:

- Not modify the evaluation pipeline.
- Not load any pretrained weights.
- Run faster than prior record when baselined on the same hardware.

Other than that, anything and everything is fair game, including changes to the synthetic prior generation!

Timing protocol

- Counting only training wall-clock time (forward/backward/optimizer steps over synthetic batches).
- Evaluation runtime is excluded.
 - the training timer is stopped while running validation.
- Prior generation time is excluded.
 - using a pre-generated prior dump is allowed.

Evaluation details

Evaluation is on all 38 TabArena classification datasets:

- if >100 features, randomly select 100
- if >1000 rows, randomly select 1000 (stratified by class labels)
- 5-fold StratifiedKFold with shuffling, class labels are encoded with integers per fold
- average binary or one-vs-rest ROC AUC over all datasets

Improvement techniques

This improvement in training speed has been brought about by the following techniques:

- Muon optimizer [3]
- Batched Muon zeropower update for grouped QKV matrices
- Scaled Dot-Product Attention rewrite with explicit QKV
- Pre-norm transformer blocks [4]
- Compile TransformerEncoderLayer forward
- bfloat16 autocast in training and inference
- Set float32 matmul precision to high
- Increase learning rate from 10^{-4} to 10^{-3}
- Increase embedding size from 192 to 256
- Reduce attention heads from 6 to 4

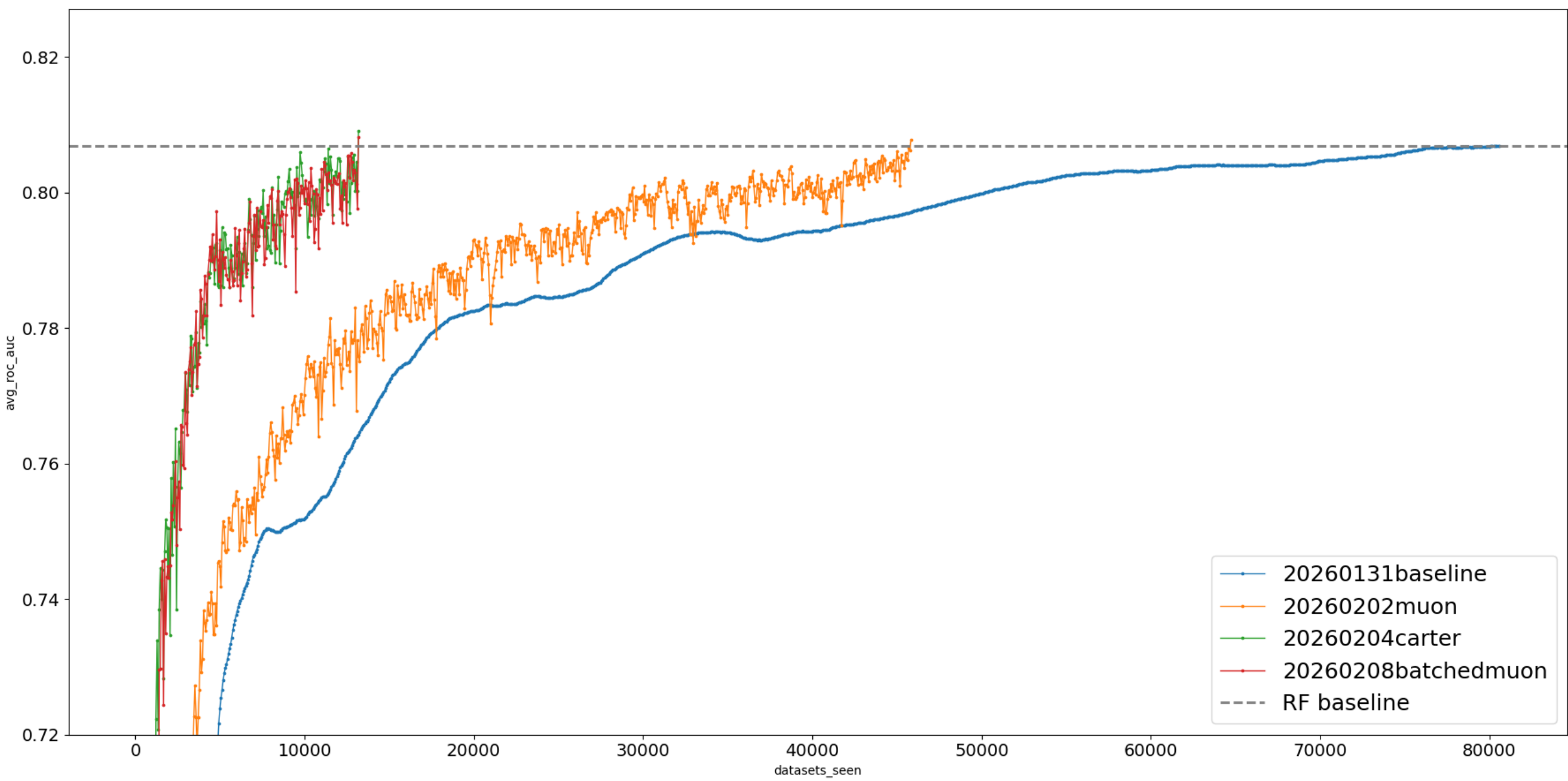
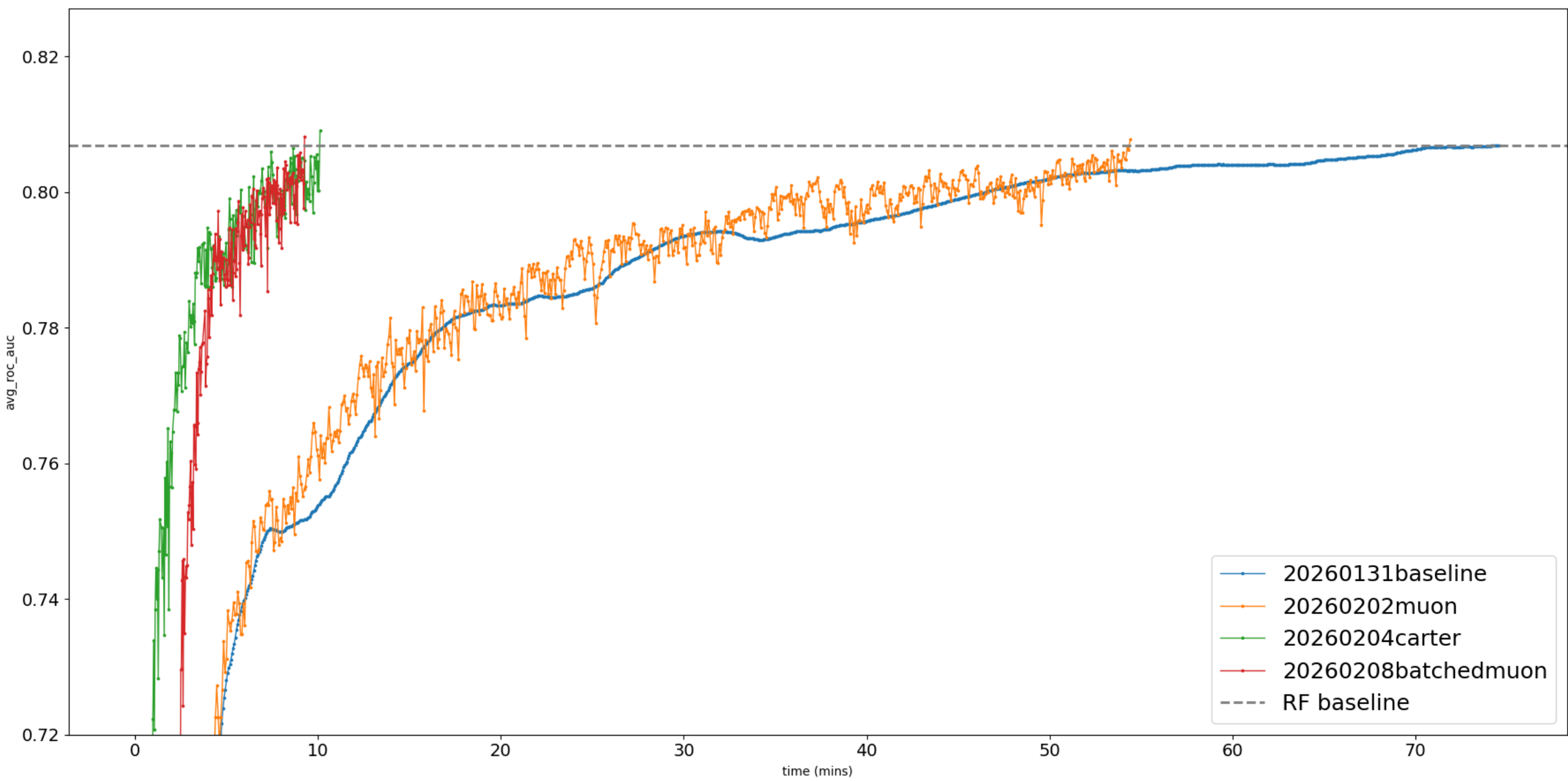
Work in progress:

- GoLU activation function [5]
- RoRa [6]

The following techniques were evaluated but did not lead to improvements:

- Initialise linear layers with Xavier initialization [7]
- Disable all linear layer biases [4]

Record history



Record time	Description	Contributors
74.32 mins	Baseline	@borawhocodess, nanotabpfn contributors
54.41 mins	Muon optimizer	@borawhocodess
10.10 mins	SDPA, bf16, higher LR, wider embeddings, fewer heads	@carterprince
9.26 mins	Batched Muon, compiled forward	@carterprince

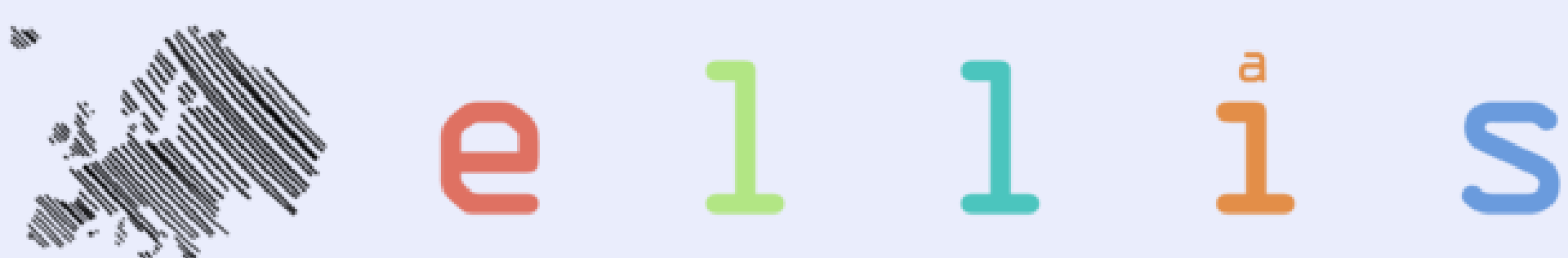
Join the race!



Scan to join the speedrun on GitHub
github.com/borawhocodess/modded-nanotabpfn

References

- [1] Noah Hollmann, Samuel Müller, Katharina Eggenberger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations (ICLR)*, 2023.
- [2] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirmer, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.
- [3] Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024.
- [4] Jonas Geiping and Tom Goldstein. Cramming: Training a language model on a single gpu in one day, 2022.
- [5] Indrashis Das, Mahmoud Safari, Steven Adriaens, and Frank Hutter. Gompertz linear units: Leveraging asymmetry for enhanced learning dynamics. *arXiv preprint arXiv:2502.03654*, 2025.
- [6] ... Rora-tab: Geometric subspace selection for tabular learning via rotational rank adaptation. 2024.
- [7] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterton, editors, *Proceedings of the Thirtieth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.
- [8] Keller Jordan, Jeremy Bernstein, Brendan Rappazzo, @fernbear.bsky.social, Boza Vlado, You Jiacheng, Franz Cesista, Braden Koszarsky, and @Grad62304977. modded-nanogpt: Speedrunning the nanogpt baseline, 2024.
- [9] Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josef Grabocka, and Frank Hutter. Transformers can do Bayesian inference. In *International Conference on Learning Representations (ICLR)*, 2022.
- [10] Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Mutalik Desai, David Salinas, and Frank Hutter. Tabarena: A living benchmark for machine learning on tabular data, 2025.
- [11] Alexander Pfefferle, Johannes Hog, Lennart Purucker, and Frank Hutter. nanotabpfn: A lightweight and educational reimplement of tabpfn, 2025.



INSTITUTE
TÜBINGEN

