Notes on Performance Prediction for Large Systems via Text-to-Text Regression

compiled by D. Gueorguiev, Intro by David Sauerwein, 8/31/25

# Introductory words

Tabular data has long resisted the deep learning revolution - but no more. And TabArena is the first living benchmark to systematically track this revolution. Here is a summary of how it works:

The incredible progress in computer vision and NLP over the last decade can be traced to three key properties of the corresponding data:

1) Strong inherent structure: The manifold hypothesis holds true for images and text, allowing extremely efficient descriptions of high-dimensional data.

2) High-quality, redundant data: For virtually all images and texts available, humans can quickly extract their essence and meaning.

3) Massive scale: Incredible amounts of internet data enable neural networks to uncover structure through gradient descent.

None of these properties apply to tabular data.

Tabular datasets are heterogeneous, messy, and often small—conditions where tree-based algorithms continued to dominate practical ML.

But new architectures (e.g. transformers), increased compute, and novel data techniques have enabled deep learning models to gradually crack the tabular space.

However, researchers claiming their DL models "crushed" traditional ML were repeatedly proven wrong. Their benchmarks were often outdated or biased, with new algorithms ultimately beaten by out-of-the-box XGBoost.

TabArena (link in comments) solves this benchmarking problem with the first living benchmark. It's treating evaluation as versioned software that's professionally maintained and community-improved.

Here are 4 key features:

------------------------

1) 1,053 datasets from tabular research, with 51 representing manually curated real-world tasks

2) 16 models including 3 tabular foundation models maintained by the team and tested in production-grade pipelines.

3) Public leaderboard with precomputed results and reproducible code to accelerate comparison.

4) A Maintainer team from multiple institutions experienced in maintaining open-source projects.

The results are really exciting:

------------------------------

(A) Best individual model performance typically comes from post-hoc ensembling tuned hyperparameters configurations

(B) With proper tuning and ensembling, deep learning methods equal or exceed gradient-boosted trees

(C) Tabular foundation models dominate small data scenarios through strong in-context learning

(D) Ensemble pipelines are state-of-the-art, though individual models contribute unequally

For the first time, we can definitively track whether and where deep learning has truly conquered tabular data. An important step to tap into the full potential of all these millions of underutilized SQL tables and Excel files lying around in enterprises arounds the globe.

A graph of different colored bars

AI-generated content may be incorrect.

References

[1] [Performance Prediction for Large Systems via Text-to-Text Regression, Y. Akhauri et al, Google, Cornell U., NCSU, 2025](https://github.com/dimitarpg13/large_language_models/blob/main/articles/tabular_data/Performance_Prediction_for_Large_Systems_via_Text-to-Text_Regression_Akhauri_2025.pdf)

[2] <https://github.com/google-deepmind/regress-lm>

[3] [Simulating large systems with Regression Language Models, Y. Akhauri et al, Google, 2025](https://research.google/blog/simulating-large-systems-with-regression-language-models/)