Beyond IID: How General Are Tabular Foundation Models, Really?

Lennart Purucker[1,2], Andrej Tschalzev[3], Nick Erickson[1], Gioia Blayer[4], David Holzmüller[4], Alan Arazi[5,1], Alexander Pfefferle[6,2], Mustafa Tajjar[2,7], Gaël Varoquaux[4,8], Frank Hutter[1,6,2]

[1] Prior Labs [2] University of Freiburg [3] University of Mannheim [4] INRIA Saclay [5] Technion [6] ELLIS Institute Tübingen [7] Zuse School ELIZA [8] Probabl

PRIOR

mail@tabarena.ai

1 Summary

How do **tabular foundation models** (TFMs) generalize beyond IID?

The first **unified benchmark** for tabular data, with:

- ❖ **IID (random), temporal, and grouped** splits
- ❖ **100 to 1M** training samples; low and high dimensionality
- ❖ **Text** and **high-cardinality** categorical features

Contributions

- Curated **142 of 1128** datasets by hand, plus introduced the **Data Foundry** framework for reproducible data curation
- Evaluated ~~3~~ **18 TFMs** vs. ~~8~~ **9 strong baselines** (~~11~~ **27** models)
- Integrated into **TabArena**: live leaderboard, code, and data

2 Background

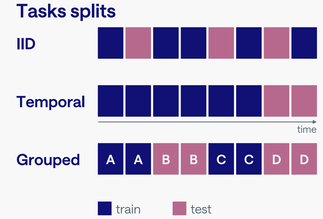
Tabular data

num	cat	text	date	target
3.2	red	"late"	2024-05	1
0.7	blue	"on time"	2024-06	0
5.1	red	"refund"	2024-07	1
2.4	green	"paid"	2024-08	?

features

Rows are samples, **columns** are features; predict the **target**.

Tasks splits



Temporal: test after train. **Grouped**: unseen groups.

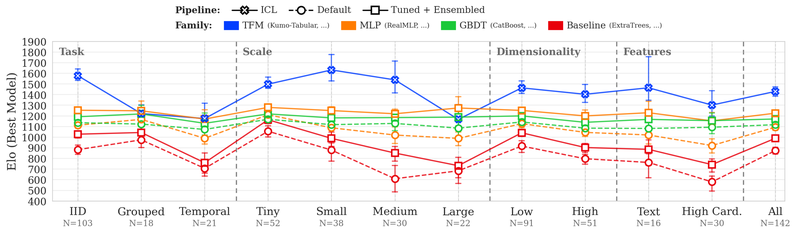
3 Related tabular benchmarks

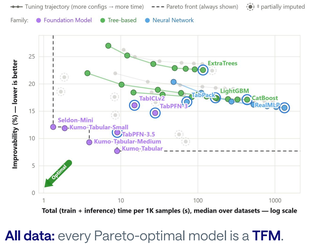
Benchmark	IID	Temp.	Group.	Text	Tiny	Large	Living
TabArena Erickson et al., 2025	●						●
TALENT Liu et al., 2024	●					●	
PMLB-Mini Knauer et al., 2024	●				●		
TabReD Rubachev et al., 2024		●				●	
TextTabBench Mráz et al., 2025	●			●			
TableShift Gardner et al., 2023			●			●	
BeyondArena ours	●	●	●	●	●	●	●

4 Benchmark design

- **Models:** tabular foundation models, MLPs, GBDTs, and simple baselines
- **Implementation: TabArena** / **AutoGluon** models, unit-tested and **open source**
- **Outer splits: repeated CV** (3 to 20 repeats by size), **grouped CV**, or **application-specific temporal splits**
- **Tuning: 8-fold inner CV** (5 × 5 below 500 samples), **non-IID** when needed; **25 random configs** plus **post-hoc ensembling**; TFMs **in-context only**
- **Preprocessing:** dates via **skrub**; text via **Qwen3-Embedding-8B**; and group index **dropped or encoded**
- **Metrics: ROC AUC**, **log-loss**, **RMSE**; aggregated with **Elo** and **improvability**
- **Cost:** compute costs for the full benchmark on cloud compute, **roughly 50k$**

5 BeyondArena live leaderboard (17th of November 2026)

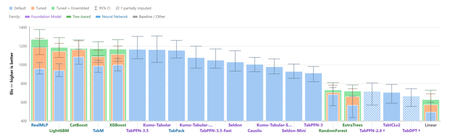




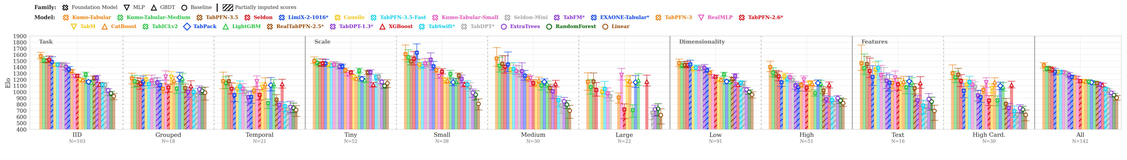
All data: every Pareto-optimal model is a **TFM**.

Key takeaways from the live leaderboard

- **TFMs lead** on IID, tiny-to-medium, text, and high-cardinality data.
- **Grouped and large data stay open:** tuned + ensembled MLPs and GBDTs match or beat TFMs.



Large data (>100k rows): best TFM ranks only **6th**.



6 Data curation

(0) Dataset Discovery

1128 datasets from **21** benchmarks and public data repositories

(1) No duplicates

-369 datasets
759 remain

(2) No few-shot prediction task

-11 datasets
748 remain

(3) Must be from a predictive task

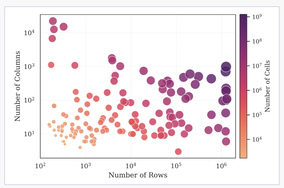
-196 datasets
552 remain

(4) Represents a real application

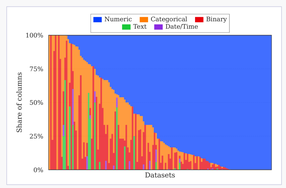
-171 unsuited modality
-115 artificial datasets
-89 data quality issues
-29 trivial tasks
-6 ethical concerns

-410 datasets
142 remain

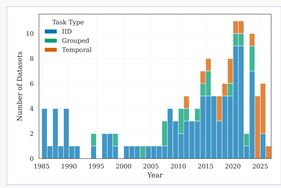
142 tiny to large-sized, tabular IID and non-IID tasks



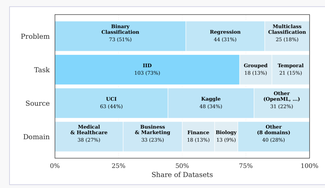
Dataset size: **rows × columns**, coloured by cells.



Column types per dataset.



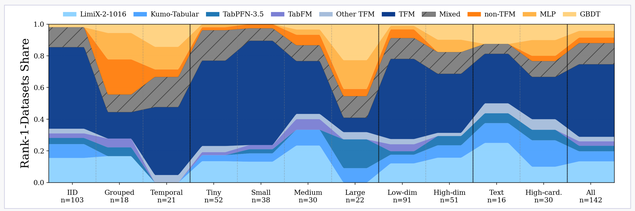
Dataset year by task type: mostly recent.



Problem, task, source, and **domain** shares.

7 Peak performance: TFMs vs rest

88% of datasets where a TFM reaches **peak performance**. The best TFMs rank first on **46%** and **40%**.



Share of datasets where a model (family) **significantly wins**.

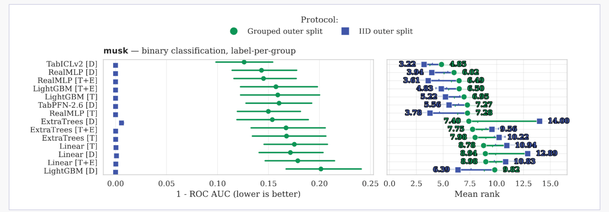
8 Scale: ≤100k vs >100k

- **≤100k: TFMs lead** on tiny, small, and IID data
- **>100k:** tuned + ensembled **GBDTs and MLPs win**

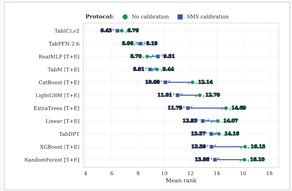
From the paper · Fig. 1 + 4, 11 models, June 2026

Full paper and results: scan the **Paper QR** below.

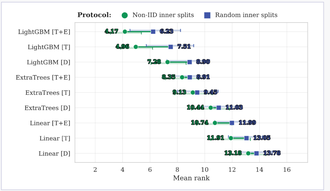
9 Insights from ablations



IID outer splits on grouped data **distort rankings** (musk shown; τ = 0.49 and 0.60).



Calibration improves ranks for most models, except TabPFN-2.6 and RealMLP.



Non-IID inner splits beat random inner splits on non-IID tasks.

Paper
arXiv:2606.30410

Code
tabarena.ai/code

DataFoundry
github.com/TabArena/data-foundry

Careers
priorlabs.ai/careers

PRIOR · ufr · Université de Fribourg · Inria · TECHNION Israel Institute of Technology · ellis Institute Tübingen · ZUSE SCHOOL ELIZA · :probabl. · small data MORE FROM LESS CRC 1597