





STRABLE: Benchmarking Tabular Machine Learning with Strings

Gioia Blayer, Myung Jun Kim, Félix Lefebvre, Lennart Purucker, Alan Arazi, Eilam Shapira, Roi Reichart, Frank Hutter, Marine Le Morvan, David Holzmüller, Gaël Varoquaux

SODA Team, Inria Saclay · Probabl · University of Freiburg · Prior Labs · ELLIS Institute Tübingen · Technion

Strings carry signal that numbers miss

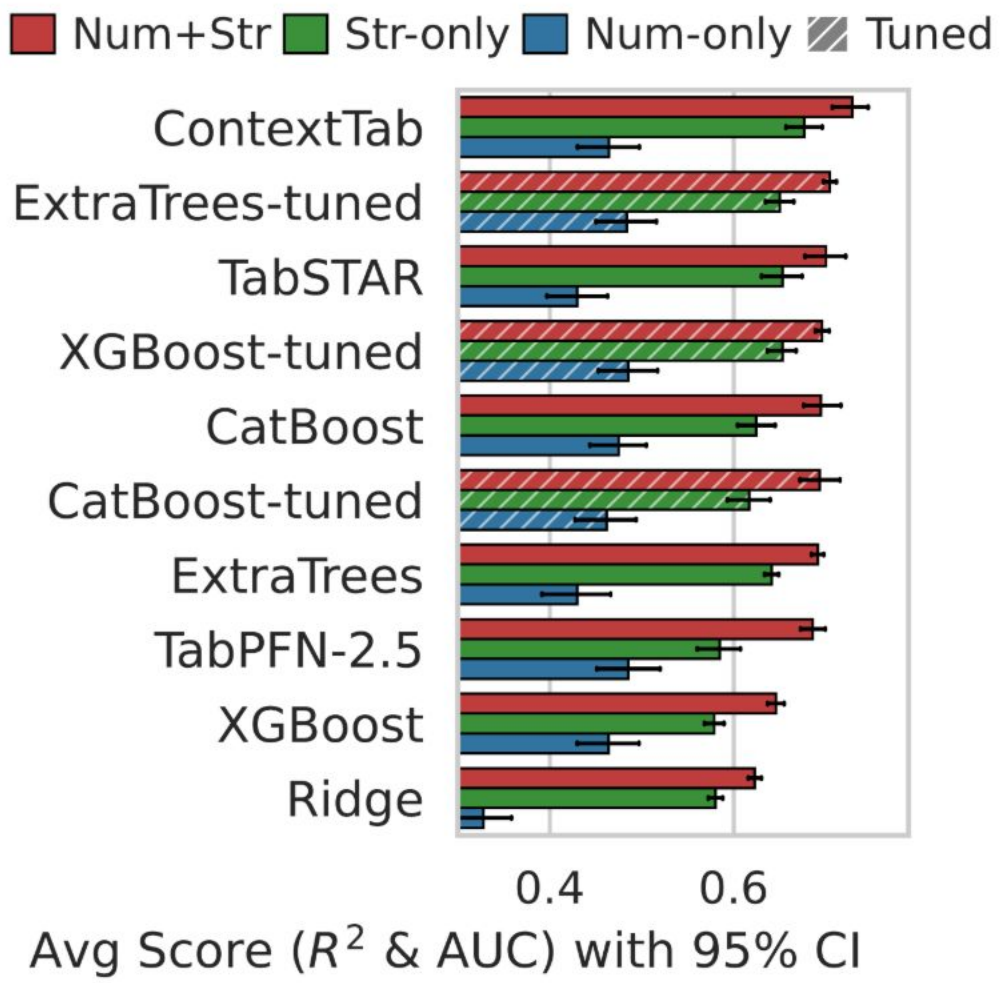
How should strings enter a tabular model?

Simple encoders
Tf-Idf, TargetEncoder

Pre-trained language models
FastText, MiniLM … Qwen-3-8B

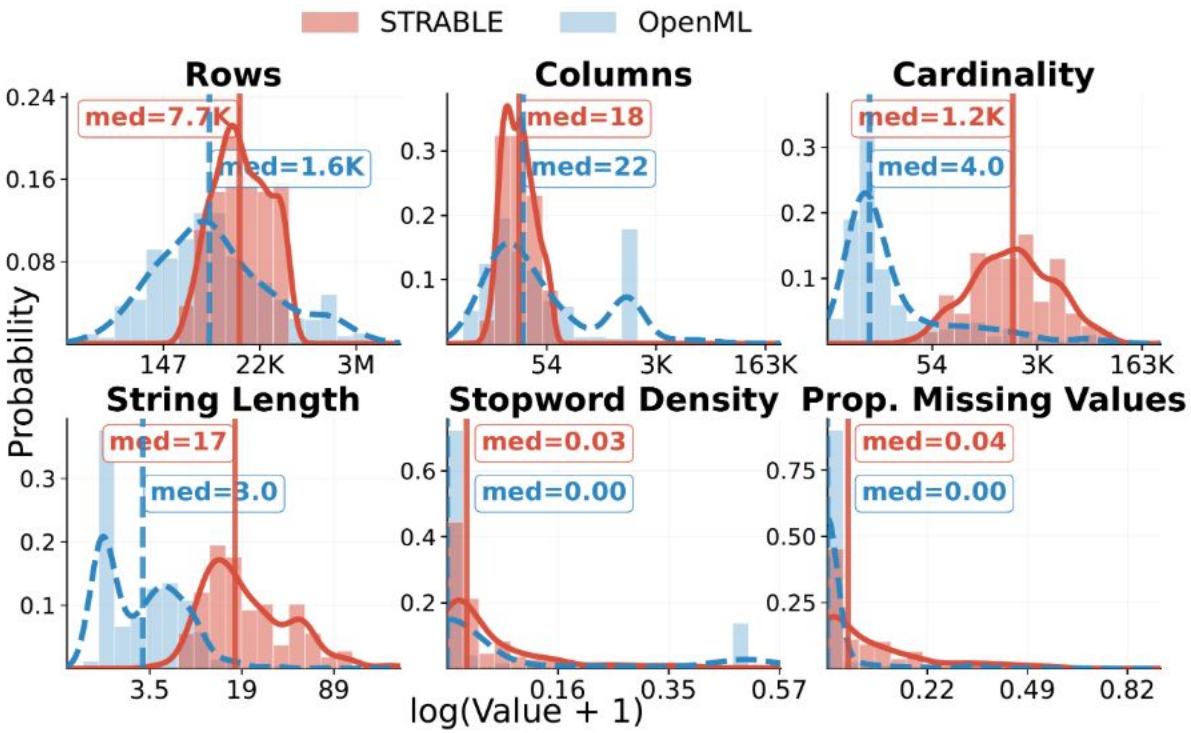
End-to-end models
CatBoost, Mambular, TabSTAR, ContextTab

STRABLE: the first large-scale empirical comparison



The STRABLE corpus

108	33	8	445
tables	sources	application fields	pipelines



Median	STRABLE	OpenML
Unique values	1.2K	4
String length	17	3
Rows	7.7K	1.6K

76 regression, 32 classification tasks, from open repositories and community platforms

Tables are left **minimally preprocessed** on purpose

Only 8% of string columns are free text



49%
Categorical
e.g. “Full-time”

23%
Names
e.g. “Acme Corp”

17%
Structured codes
e.g. “AB-1042”

8%
Free text
e.g. “Great food, but the service…”

2%
Datetime
e.g. “2024-03-01”

0.5%
Identifiers
e.g. “id_83921”

A heuristic on simple column indices assigns the type: **97% agreement** on 30 hand-labelled tables. Examples are illustrative.

Two designs: modular and end-to-end



vs. End-to-end Raw table in, one model: CatBoost, Mambular, TabSTAR, ContextTab

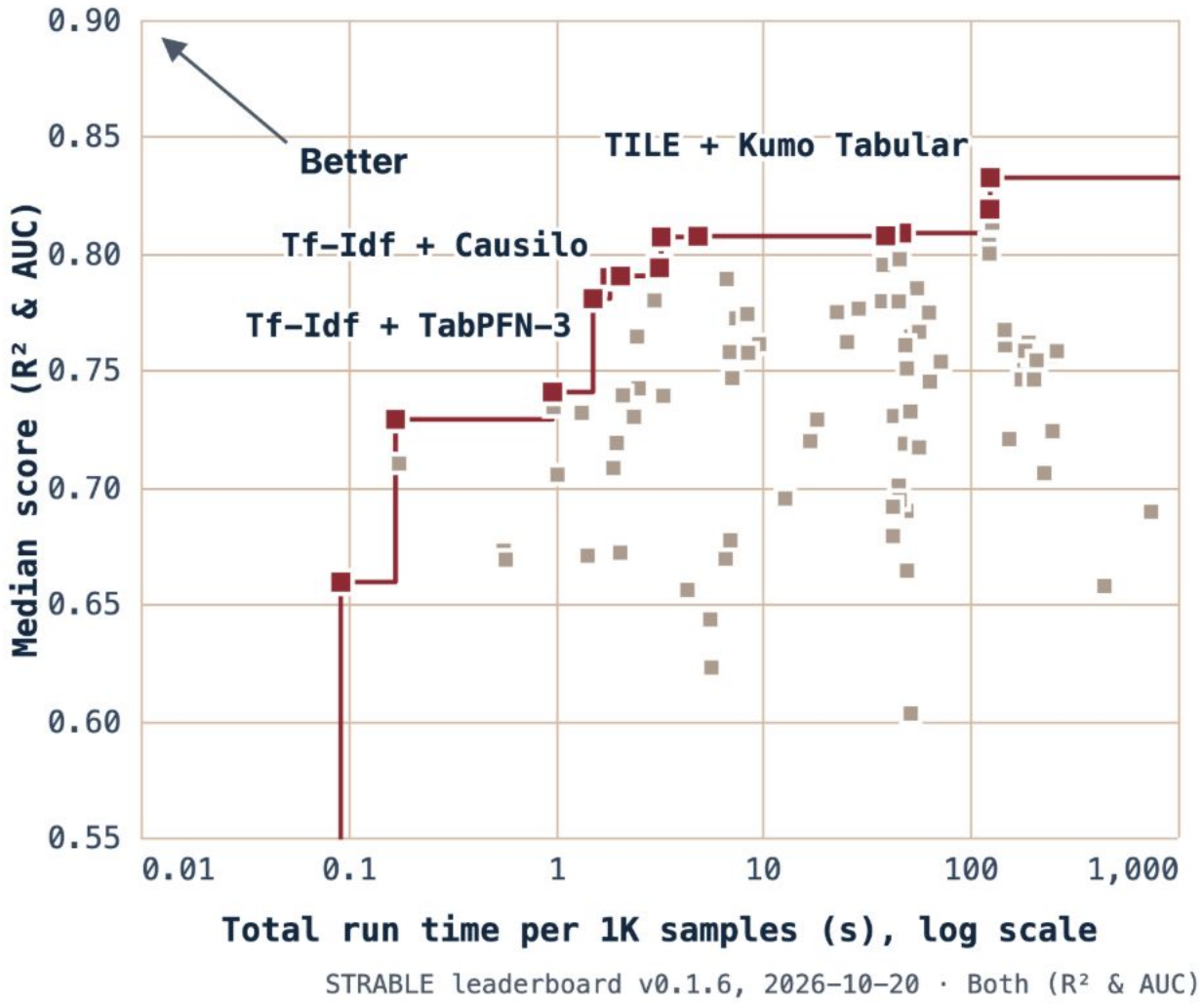
Each cell is encoded independently. LM embeddings use **mean pooling** over the last layer, reduced per text feature with **30-PCA** by default.

Modular beats end-to-end on the Pareto front

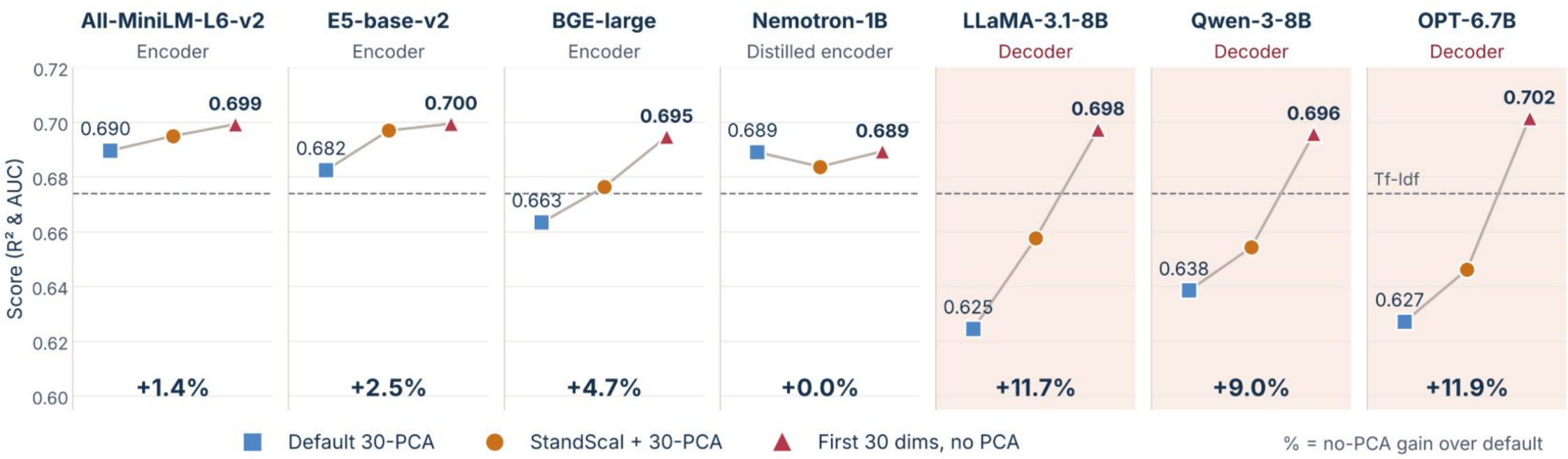
Encoder + post-processing + learner pipelines **dominate the frontier** of score and run time.

The leaderboard **keeps growing** as new models are added, e.g. TILE + Kumo Tabular.

gblayer.github.io/ strable-leaderboard



LLM embeddings need careful post-processing



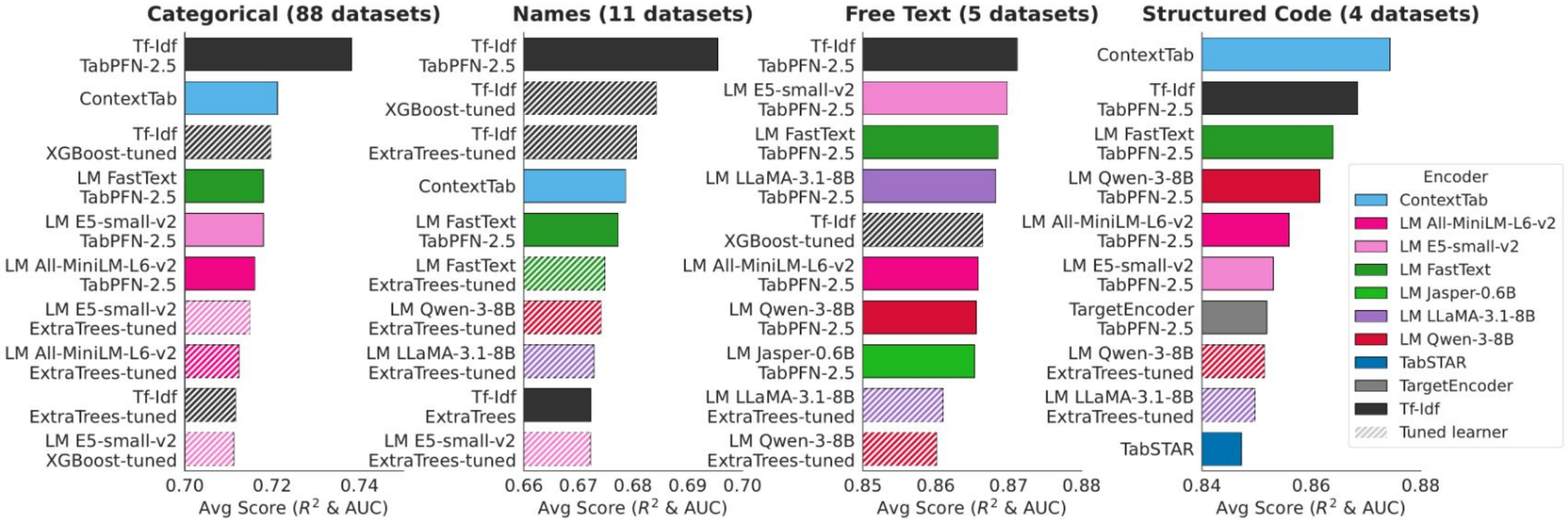
Depends on the language model

Plain PCA suits encoders but hurts decoder-only LLMs. Scaling first, or keeping the first 30 dims, recovers up to **+11.9%**.

In decoders, a few dimensions dominate

Gini of per-dim variance (higher = fewer dims dominate): LLaMA **0.42**, OPT **0.36**, Qwen-3 **0.24**, vs **0.11–0.15** for encoders.

Free text is where LLMs gain ground



Most tables

Most strings are short and repetitive, so a simple encoder is enough.

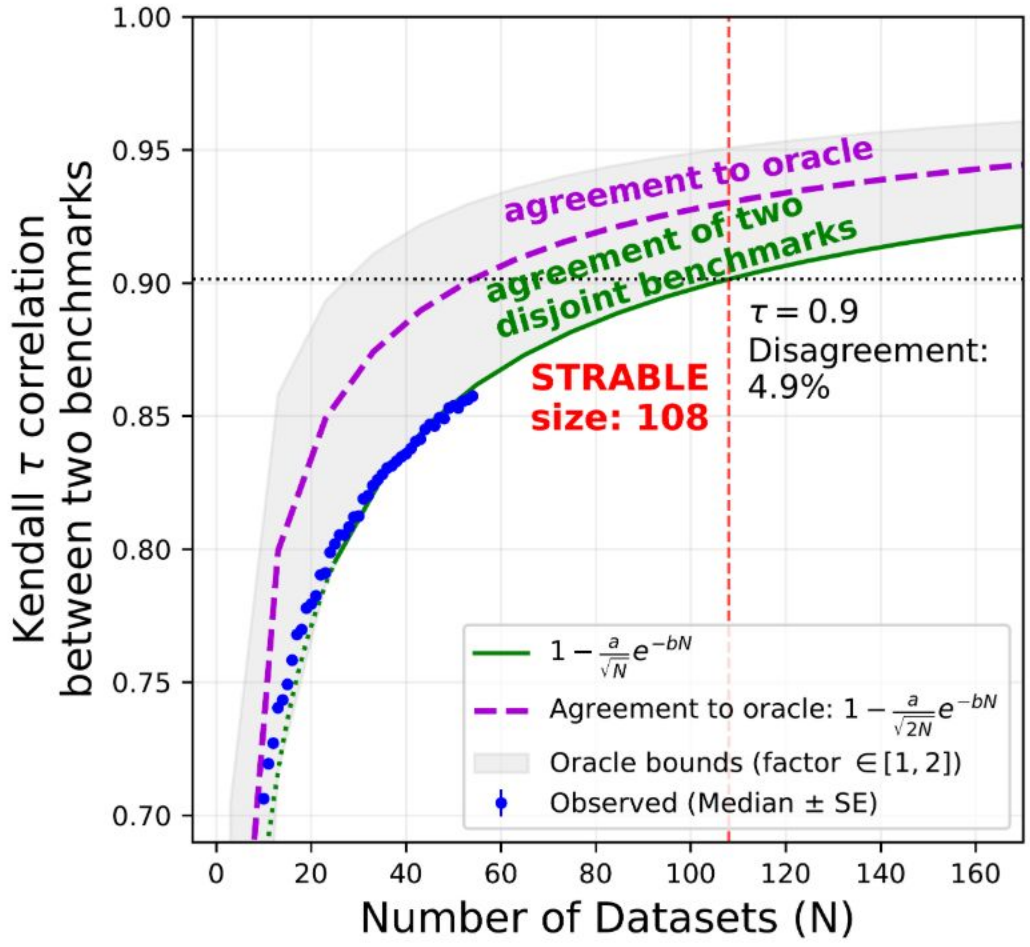
Free-text tables

For tables whose free text is the leading string type, every large LLM enters the **top 10**.

Words per cell

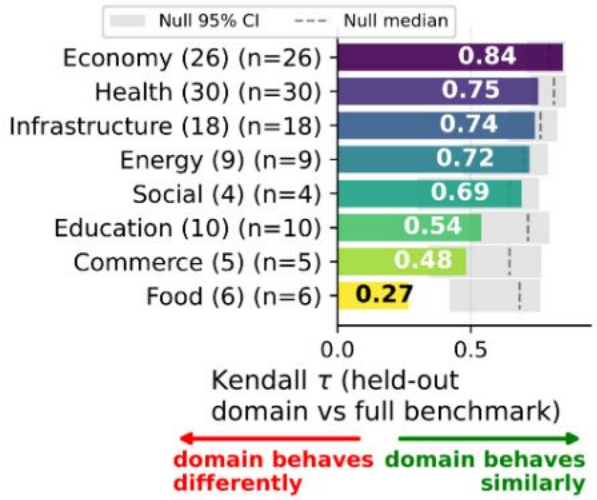
Strongest ranking disruptor: $\tau \approx 0.09$, short vs long text

STRABLE rankings generalize



At 108 tables: τ ≈ 0.90 between two benchmarks, a lower bound, and **τ ≈ 0.93 to the oracle**

Across application fields (Kendall τ)



Across data preparation (Kendall τ)

Subsampling	0.96
Target transform	0.95
Imputation	0.83
Feature eng.	0.70

Takeaways

01 **Modular pipelines** give the best trade-off of performance and runtime

02 LLM encoders pay off on **longer free text**

03 Results are sensitive to **how embeddings are extracted**

Project page

Leaderboard

gioia.blayer@inria.fr