A Tale of Two Audits

Ratish Puduppully

24 April 2026

IT University of Copenhagen

Audit 1 — The Illusion of Generalization in Tabular Language Models

Aditya Gorla (UCLA) & Ratish Puduppully

Under review at ICML 2026

Why tabular data?

A lot of real-world data is tabular — healthcare, finance, government, science.

age	sex	BP	HR	notes	**readmit?**
67	M	140/90	85	*"chest pain, SOB on exertion"*	yes
45	F	120/80	72	*"routine follow-up, stable"*	no
72	M	160/95	95	*"post-op wound infection, fever"*	yes

Foundation models transformed text and vision. **Not** tabular prediction — **classical (non-neural) methods** still win.

Tabular Language Models (TLMs) claim: do for tables what LLMs did for text.

What is a TLM?

age	sex	BP	HR	notes	**readmit?**
67	M	140/90	85	*"chest pain, SOB on exertion"*	**?**

What is a TLM?

age	sex	BP	HR	notes	**readmit?**
67	M	140/90	85	*"chest pain, SOB on exertion"*	**?**

↓ serialize

```
Predict the value of readmit:  ||yes||no||
The age is 67.  The sex is M. The BP is 140/90.  The HR is 85.
The notes is ‘‘chest pain, SOB on exertion’’.
What is the value of readmit?  ||yes||no||
```

What is a TLM?

age	sex	BP	HR	notes	**readmit?**
67	M	140/90	85	*"chest pain, SOB on exertion"*	**?**

↓ serialize

```
Predict the value of readmit:  ||yes||no||
The age is 67.  The sex is M. The BP is 140/90.  The HR is 85.
The notes is ‘‘chest pain, SOB on exertion’’.
What is the value of readmit?  ||yes||no||
```

↓ LLM

yes

What is a TLM?

age	sex	BP	HR	notes	**readmit?**
67	M	140/90	85	*"chest pain, SOB on exertion"*	**?**

↓ serialize

```
Predict the value of readmit:  ||yes||no||
The age is 67.  The sex is M. The BP is 140/90.  The HR is 85.
The notes is ‘‘chest pain, SOB on exertion’’.
What is the value of readmit?  ||yes||no||
```

↓ LLM

`yes`

TLM: fine-tune an LLM on millions of such prompts; claim zero-shot on new datasets.

The Tabula-8B paper

Large Scale Transfer Learning for Tabular Data via Language Modeling

Josh Gardner[1,*] **Juan C. Perdomo**[2] **Ludwig Schmidt**[1,3]

[1]**University of Washington**, [2]**Harvard University**, [3]**Stanford University**
[*]Corresponding author, jpgard@cs.washington.edu

Abstract

Tabular data – structured, heterogeneous, spreadsheet-style data with rows and columns – is widely used in practice across many domains. However, while recent foundation models have reduced the need for developing task-specific datasets and predictors in domains such as language modeling and computer vision, this transfer learning paradigm has not had similar impact in the tabular domain. In this work, we seek to narrow this gap and present TABULA-8B, a language model for tabular prediction. We define a process for extracting a large, high-quality training dataset from the TabLib corpus, proposing methods for tabular data filtering and quality control. Using the resulting dataset, which comprises over 2.1B rows from 4.2M unique tables, we fine-tune a Llama 3-8B large language model (LLM) for tabular data prediction (classification and binned regression) using a novel packing and attention scheme for tabular prediction. Through evaluation across a test suite of 329 datasets, we find that TABULA-8B has zero-shot accuracy on unseen tables that is over 15 percentage points (pp) higher than random guessing, a feat that is not possible with existing state-of-the-art tabular prediction models (e.g. XGBoost, TabPFN). In the few-shot setting (1-32 shots), without any fine-tuning on the target datasets, TABULA-8B is 5-15 pp more accurate than XGBoost and TabPFN models that are explicitly trained on equal, or even up to 16× more data. We release our model, code, and data along with the publication of this paper.[†]

1 Introduction

Transfer learning - the ability of a model to accurately solve prediction tasks on data it was not trained on - is one of the defining hallmarks of recent foundation models in domains such as vision [38] and language [6]. Among their many advantages, transferable models expand the scope of problems that can be tackled via machine learning by reducing the need for curated, task-specific models and datasets. Such models also can provide both absolute performance and sample-efficiency gains over task-specific models when applied to new tasks [38, 41, 53].

In this work, we introduce a new model and dataset for large-scale transfer learning on tabular data. Tabular, spreadsheet-style data underlies applications in healthcare, finance, government, and the natural sciences [4, 16, 58].

[†]For links to all code, data, and model, see Section 7.

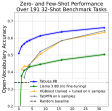

Figure 1: TABULA-8B outperforms SOTA tabular baselines across 0 – 32-shot tasks from five tabular benchmarks.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Our re-evaluation

We re-ran Tabula-8B on its own **165-dataset benchmark**:

- 64 binary classification
- 55 multiclass classification
- 46 quartile — regression targets binned into 4

Question: Is the reported performance generalization — or something else?

Binary example: hospital readmission

age	sex	BP	HR	notes	**readmit?**
67	M	140/90	85	*"chest pain, SOB on exertion"*	yes
45	F	120/80	72	*"routine follow-up, stable"*	no
72	M	160/95	95	*"post-op wound infection, fever"*	yes

Predict: will this patient be readmitted within 30 days?

Multiclass example: Wine Quality

volatile ac.	sugar	pH	sulphates	alcohol	**quality (3–8)**
0.70	1.9	3.51	0.56	9.4	5
0.88	2.6	3.20	0.68	9.8	5
0.28	1.9	3.16	0.58	9.8	6

Predict: wine quality score from chemistry.

Quartile example: California Housing

MedInc	HouseAge	AveRooms	Lat	Long	**MedHouseVal bin**
8.33	41	6.98	37.88	-122.23	*greater than 2.65*
3.20	35	4.80	36.50	-119.80	*between 1.20 and 1.80*
1.65	28	3.50	36.05	-119.30	*less than 1.20*

Predict: which bin the median house value falls in.

Finding 1 — The missing baseline

Dataset	Accuracy
Brain Stroke	95.8%
All Space Missions	90.4%
Uber Data	84.5%
Bank Personal Loan	84.3%

Finding 1 — The missing baseline

Dataset	Accuracy	Majority
Brain Stroke	95.8%	95.9%
All Space Missions	90.4%	90.3%
Uber Data	84.5%	93.2%
Bank Personal Loan	84.3%	89.9%

Finding 1 — The missing baseline

Dataset	Accuracy	Majority	Lift
Brain Stroke	95.8%	95.9%	−0.1%
All Space Missions	90.4%	90.3%	+0.1%
Uber Data	84.5%	93.2%	−8.7%
Bank Personal Loan	84.3%	89.9%	−5.7%

65 / 165 datasets (**39%**) $\leq$ majority baseline.

Finding 2a — Contamination: complete train–test overlap

Test row 728

```
kids618 = 3
age = 41
nwifeinc = 77.0
lfp = ?
```

Finding 2a — Contamination: complete train–test overlap

Test row 728

```
kids618 = 3
age = 41
nwifeinc = 77.0
lfp = ?
```

Training data match

```
k618 = 3
age = 41
inc = 77.0
lfp = no
```

Finding 2a — Contamination: complete train–test overlap

Test row 728

```
kids618 = 3
age = 41
nwifeinc = 77.0
lfp = ?
```

Training data match

```
k618 = 3
age = 41
inc = 77.0
lfp = no
```

`us-womens-labor`: **753 / 753** test rows in training data with labels.

99.6% accuracy → memorization, not generalization.

Finding 2b — Contamination: tasks leak across tables

peloton-data: Date → Day

```
Date = 11/30/2021
Instructor = Ally Love
Day = ?
```

Finding 2b — Contamination: tasks leak across tables

peloton-data: Date → Day

```
Date = 11/30/2021
Instructor = Ally Love
Day = ?
```

Training table: toronto-transit-delays

```
Date:  2021-11-30
Day:   Tuesday
Line:  501, ...
```

Finding 2b — Contamination: tasks leak across tables

peloton-data: Date → Day

```
Date = 11/30/2021
Instructor = Ally Love
Day = ?
```

Training table: toronto-transit-delays

```
Date:  2021-11-30
Day:   Tuesday
Line:  501, ...
```

Multiple date→day matches across unrelated tables —
the task is memorized even though no row is.

Finding 3 — Instruction-tuning closes the gap

Alpaca = Llama-3-8B + 50K general instructions, *zero tabular data*.

Binary + categorical classification (119 datasets)

Model	**Acc.**	**Lift**	**Δ vs Tabula**
Base Llama	47.7	−9.4	−15.8

Finding 3 — Instruction-tuning closes the gap

Alpaca = Llama-3-8B + 50K general instructions, *zero tabular data*.

Binary + categorical classification (119 datasets)

Model	Acc.	Lift	Δ vs Tabula
Base Llama	47.7	−9.4	−15.8
Tabula-8B	63.5	+6.5	—

Finding 3 — Instruction-tuning closes the gap

Alpaca = Llama-3-8B + 50K general instructions, *zero tabular data*.

Binary + categorical classification (119 datasets)

Model	Acc.	Lift	Δ vs Tabula
Base Llama	47.7	−9.4	−15.8
Tabula-8B	63.5	+6.5	—
Alpaca	**58.6**	+1.6	**−4.9**

Audit 1: summary

1. **The missing baseline** — 39% of datasets: zero or negative lift over majority
2. **Contamination in top performers** — complete overlap + task leakage
3. **Instruction-tuning alone recovers most performance**

Format adaptation + memorization, not tabular reasoning.

Audit 2 — FineWeb-Edu Misinformation Audit

Ratish Puduppully

huggingface.co/datasets/ratishsp/fineweb-edu-misinfo

Ongoing work

FineWeb and FineWeb-Edu

FineWeb (Penedo et al., NeurIPS 2024 — HuggingFace)
15 trillion tokens from 96 Common Crawl snapshots

FineWeb-Edu — subset filtered for "educational value"
1.3 trillion tokens / **1.53 billion** documents

Pretraining data for open language models:
SmolLM / SmolLM2 (HuggingFace), OLMo 2 (AI2), ...

How “educational value” is measured

1. **Llama-3-70B-Instruct** scores 460K FineWeb pages on a **0–5 scale**
2. Linear regressor distils those scores into a lightweight classifier
3. Keep pages with predicted score ≥ 2.5

Rubric rewards: **structure, coherence, academic register, textbook-like presentation**; prompted to favour grade-/middle-school knowledge.

It cannot assess factual accuracy.

Five misinformation labels (+ no misinfo)

- **Climate denial** — denies or misrepresents climate science
- **Health misinfo** — antivax, quack remedies, unproven alt-medicine cures
- **Pseudoscience** — creationism, flat earth, ancient aliens
- **Hate / extremism** — hate speech, Holocaust denial, far-right extremism
- **Conspiracy / propaganda** — QAnon, chemtrails, state-sponsored disinformation

Strict rule: flag only *active endorsement*.

Satire, critical examination, minor errors, historical documents → no misinfo.

Audit design

100K flagged-domain documents

WattsUpWithThat — climate
NaturalNews — health
ICR — pseudoscience
Stormfront — hate
InfoWars — conspiracy

100K random sample

Background rate from FineWeb-Edu

Annotation

Llama 4 Maverick

IAA with Claude Sonnet 4.6:
$\kappa = 0.86$ (misinfo vs. no misinfo)

200K documents total

4.1% of random FineWeb-Edu is misinformation

Label	Count
No misinfo	95,874
Health misinformation	2,089
Pseudoscience	1,126
Conspiracy / propaganda	476
Hate / extremism	236
Climate denial	199
Misinformation total	**4,126**

1.53B documents in FineWeb-Edu

$\times$ **4.1%**

$=$ $\sim$**63 million** documents containing misinformation

39% of flagged-domain documents are misinformation

Label	Count
No misinfo	61,147
Pseudoscience	16,179
Health misinformation	10,142
Climate denial	5,393
Conspiracy / propaganda	4,878
Hate / extremism	2,261
Misinformation total	**38,853**

38.9% misinformation rate

$\sim$10$\times$ the random baseline

Examples of Misinformation

codoh.com — Holocaust denial `edu_score: 3.75`

"Fifty years ago, on April 15, 1945, British troops liberated the Bergen-Belsen concentration camp... In fact, the dead of Bergen-Belsen were, above all, unfortunate victims of war, not deliberate policy. It can even be argued that they were as much victims of Allied as of German measures..."

icr.org — pseudoscience `edu_score: 4.91`

"Glaciers once filled Yosemite Valley almost to the top of Half Dome ... leaving behind evidence of their presence and direction of travel. But if glaciers occurred after the Genesis Flood, how did ..."

Audit 2: summary

1. $\sim$**4.1%** of random FineWeb-Edu is misinformation ($\sim$63M documents)
2. $\sim$**39%** of flagged-domain documents are misinformation
3. The "educational value" classifier optimizes for form, not truth

Thanks!

Questions?