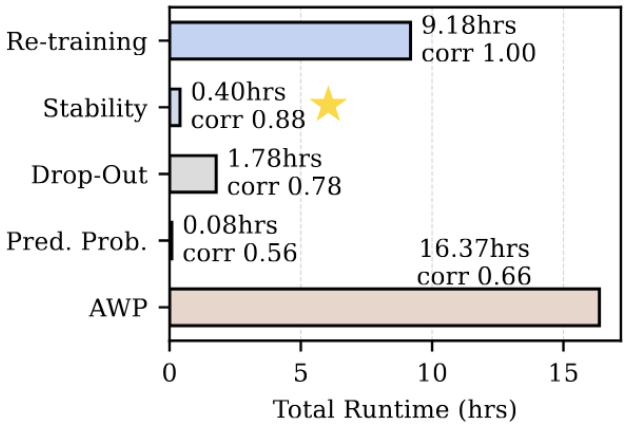
Quantifying Prediction Consistency Under Fine-Tuning Multiplicity in Tabular LLMs

Faisal Hamman[1] Pasan Dissanayake[1] Saumitra Mishra [2] Freddy Lecue[2] Sanghamitra Dutta[1]

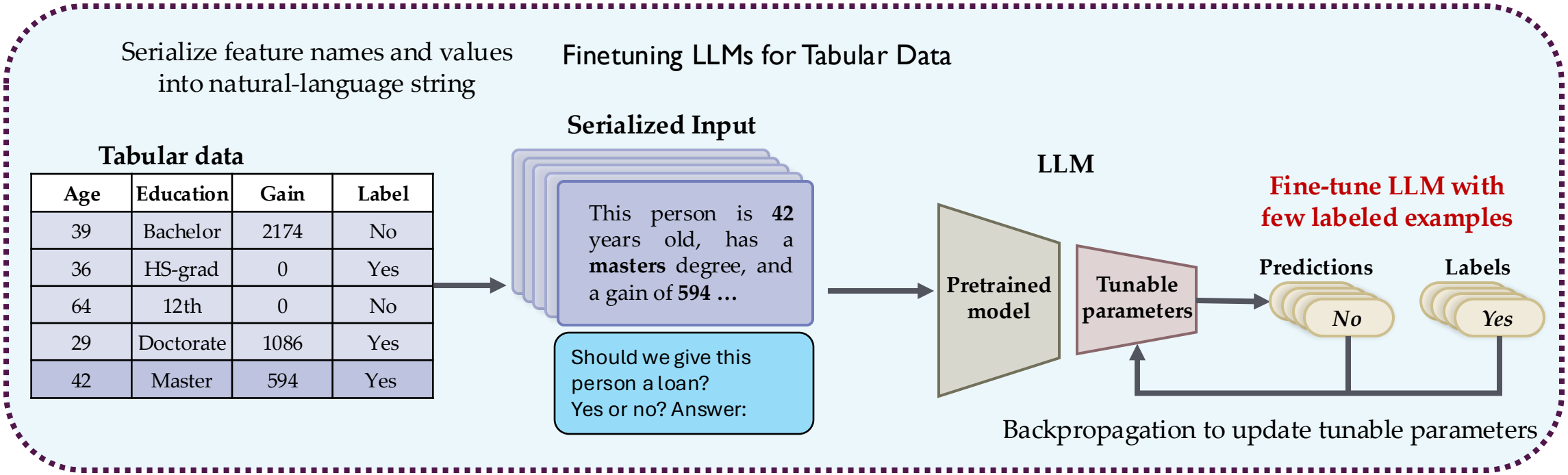
[1]University of Maryland, College Park, [2] JPMorgan AI Research



Motivation: Tabular LLMs in High-Stakes Applications

Few-shot classification on tabular datasets:

Tabular LLMs perform commendably with **very little labeled data** due to their pretrained knowledge [1,2]

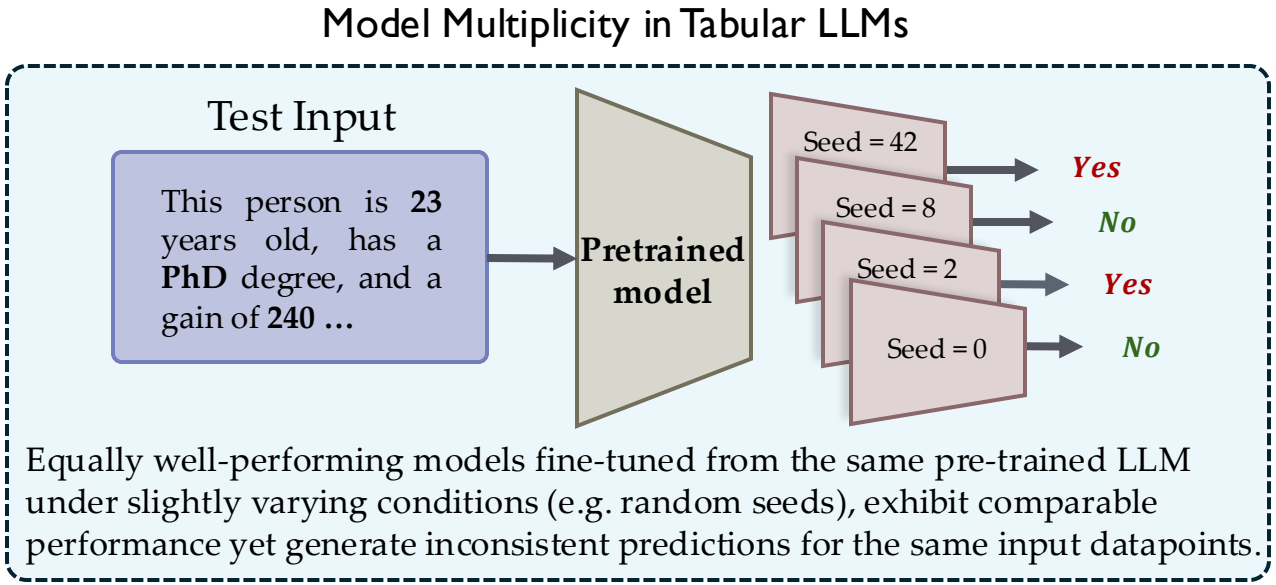


Paucity of training data + large parameter space >> **Fine-tuning multiplicity** in Tabular LLMs

[1] Hegselmann, et.al., TabLLM: Few-shot Classification of Tabular Data with Large Language Models, AISTATS 2023.

[2] van Breugel, B. and van der Schaar, M. Position: Why tabular foundation models should be a research priority, ICML 2024.

What is fine-tuning multiplicity?



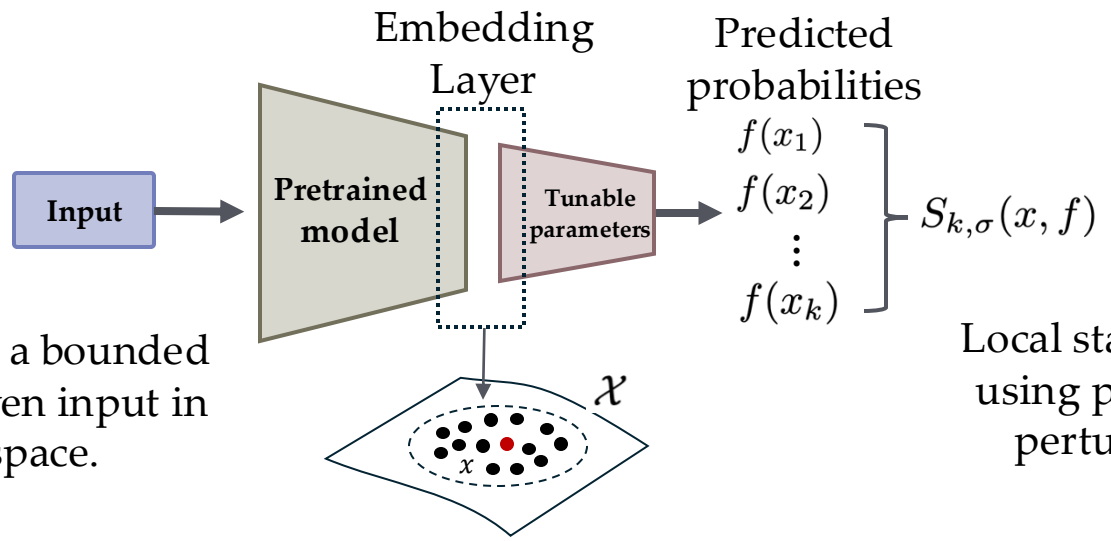
In high-stakes decision-making, arbitrary predictions can have significant consequences

Main Contributions

- Unravel the nature of **fine-tuning multiplicity** in Tabular LLMs
- A measure to quantify prediction consistency under fine-tuning multiplicity – that we call **local stability** – **does not need expensive model retraining multiple times**
- Probabilistic guarantees over a broad class of equally-well-performing fine-tuned models
- Experimental validation: **Local stability measure highly correlates with actual fine-tuning multiplicity**

Related Work: J. Gomez, C. Machado, L. Monteiro, and F. Calmon "Algorithmic Arbitrariness in Content Moderation", FAccT '24.

Our Proposed Local Stability Measure



We sample k points in a bounded neighborhood of a given input in the embedding space.

Local stability measured using predictions from perturbated points

Local Stability

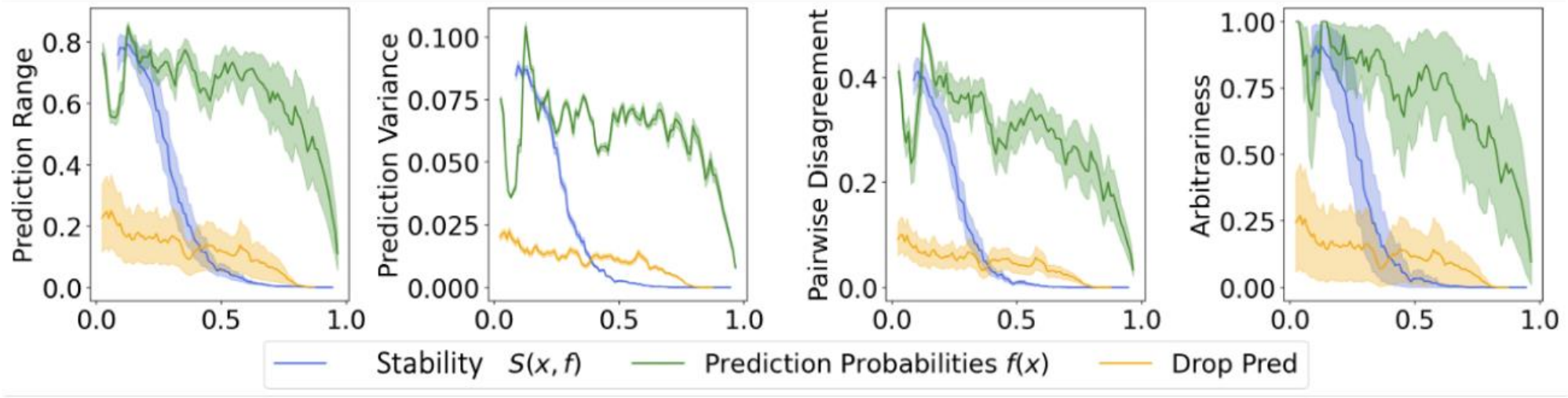
$$S_{k,\sigma}(x, f) = \frac{1}{k} \sum_{x_i \in N_{x,k}} f(x_i) - \frac{1}{k} \sum_{x_i \in N_{x,k}} |f(x) - f(x_i)|$$

$$N_{x,k} = \{x_1, x_2, \ldots, x_k\} \subset B(x, \sigma) = \{x' \in \mathcal{X} : \|x' - x\|_2 < \sigma\}.$$

Is a set of k points sampled independently from a distribution over a hypersphere of radius σ centered at x.

Probabilistic Guarantee

Informally Stated: Under mild assumptions,
datapoints with high local stability will remain **consistent** with high probability
over a broad class of equally-well-performing fine-tuned models.



Experiments

Goal: To compare our local stability measure (without retraining multiple times) with actual fine-tuning multiplicity across multiple models

Multiplicity Metrics [3]: Pairwise Disagreement, Arbitrariness, Prediction Variance and Range

Datasets: Real-world tabular datasets (e.g., German Credit, Bank, Heart, Car, Diabetes, Adult) under **few shots** (64, 128, 512).

Models: LLMs such as *Bigscience T0* and Google *Flan T5* using T-Few and LoRA.

Additional Baselines: Prediction confidence, Dropout-based methods [4], Adversarial Weight Perturbation (AWP) [5].

[3] J. Gomez, C. Machado, L. Monteiro, and F. Calmon "Algorithmic Arbitrariness in Content Moderation", FAccT '24.
[4] Hsu, H., Li, G., Hu, S., and Chen, C.-F. Dropout-based rashomon set exploration for efficient predictive multiplicity estimation. ICLR, 2024
[5] Hsu, H. and Calmon, F. Rashomon capacity: A metric for predictive multiplicity in classification. Neurips, 2022

Key Finding: Strong Correlation Observed Between our Local Stability Measure and Actual Fine-Tuning Multiplicity

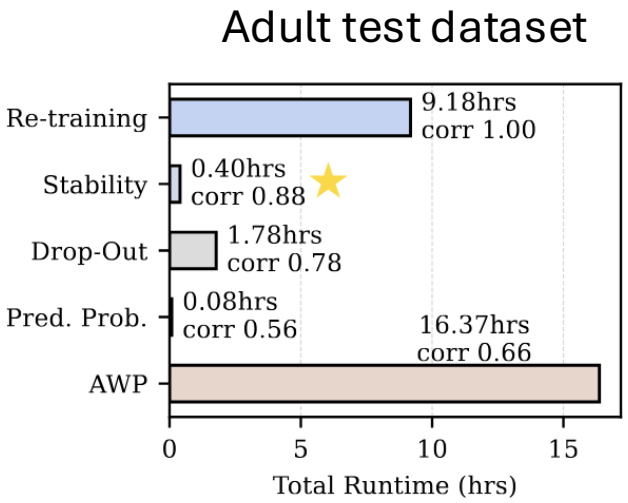
Evaluated multiplicity (assessed on 40 retrained models) versus our **local stability measure** (evaluated on one model) for the 128-shot setting.

Dataset	Number of Shots	Measure	Arbit.	Pairwise Disag.	Prediction Variance	Prediction Range
Adult	128	Pred. Prob.	0.67	0.62	0.30	0.54
		Drop-Out	0.74	0.83	0.69	0.81
		Stability	**0.80**	**0.96**	**0.84**	**0.91**
German	128	Pred. Prob.	**0.57**	**0.57**	0.86	0.86
		Drop-Out	0.50	0.56	0.74	0.84
		Stability	0.54	0.54	**0.87**	**0.87**
Diabetes	128	Pred. Prob.	0.88	0.93	**0.93**	**0.95**
		Drop-Out	0.89	0.92	0.92	0.94
		Stability	**0.92**	**0.95**	**0.93**	**0.95**
Bank	128	Pred. Prob.	0.54	0.57	0.73	0.62
		Drop-Out	0.62	0.70	0.75	0.51
		Stability	**0.79**	**0.84**	**0.87**	**0.86**
Heart	128	Pred. Prob.	0.61	0.46	0.50	0.26
		Drop-Out	0.64	0.76	0.74	0.83
		Stability	**0.89**	**0.90**	**0.97**	**0.87**
Car	128	Pred. Prob.	0.56	0.26	0.29	0.01
		Drop-Out	0.63	0.66	0.57	0.52
		Stability	**0.97**	**0.91**	**0.93**	**0.94**

- Our local stability measure $S(x, f)$ shows a higher correlation with actual multiplicity compared to baselines: prediction confidence $f(x)$ or Drop-out method.
- Our local stability measure $S(x, f)$ better informs multiplicity of a datapoint
- Additional experiments in the paper

Computational Efficiency Benefits

Total train and eval runtime across baselines.



Measure	Arbit.	Pairwise Disag.	Pred. Var.	Pred. Range	Train Time	Eval. Time
Re-training	1.00	1.00	1.00	1.00	456 mins	94.7 mins
Pred. Prob.	0.63	0.61	0.39	0.63	4.56 mins	0.51 mins
Drop-Out	0.79	0.78	0.70	0.86	4.56 mins	102 mins
AWP	0.65	0.71	0.55	0.72	4.56 mins	977.6 mins
Stability	0.81	0.96	0.80	0.93	4.56 mins	19.4 mins

Local stability measure has **significantly lower runtime** compared to the retraining and other baselines **while maintaining strong correlation** with multiplicity metrics.

Conclusion

Poster - ID 46165
Wed 16 Jul 4:30 p.m.

Thank You!

- **Novel local stability** measure to quantify prediction consistency under fine-tuning multiplicity, using **only a single model, avoiding expensive retraining (fine-tuning) multiple times.**
- **Probabilistic guarantee** showing predictions with high local stability remain consistent across a broad class of fine-tuned models with high probability.
- **Empirical Results** demonstrated that our stability measure outperforms baselines in capturing consistency, with superior correlation to fine-tuning multiplicity across various datasets.
- **Computational Efficiency**: Our method reduces complexity by avoiding retraining multiple models, requiring only inference and sampling from the embedding space.
- **Implications for Trust**: The measure helps practitioners assess which predictions to trust, reducing risks of inconsistent or conflicting outcomes in high-stakes applications.