AnoLLM: Large Language Models for Tabular Anomaly Detection

Amazon

Presenter: **Che-Ping Tsai**

Joint work with Ganyu Teng, Phil Wallis and Wei Ding

Motivation (1)

- Anomaly detection (AD) on tabular data has a wide range of applications.

Cyber-attack prevention

Fraudulent financial transaction detection

Unusual medical condition detection

Motivation (2)

- LLMs excel in NLP tasks.
- They perform well in other modalities (e.g., vision, tabular data).
- Tabular anomaly detection remains unexplored.

Motivation (2)

- LLMs excel in NLP tasks.
- They perform well in other modalities (e.g., vision, tabular data).
- Tabular anomaly detection remains unexplored.

Goal

Investigate the capabilities of LLMs in **tabular anomaly detection**.

Motivation (2)

- LLMs excel in NLP tasks.
- They perform well in other modalities (e.g., vision, tabular data).
- Tabular anomaly detection remains unexplored.

We show that **LLMs outperform existing AD methods** when data contain **mixed-typed features**!

Challenges in applying LLMs to tabular AD(1)

- Tabular data is inherently **structured**.
- However, LLMs only take sequential inputs.

Age	Accident Area	Deductible	Vehicle Price
21	Urban	300	More than 69,000
⋮	⋮	⋮	⋮
34	Rural	200	20,000 to 29,000

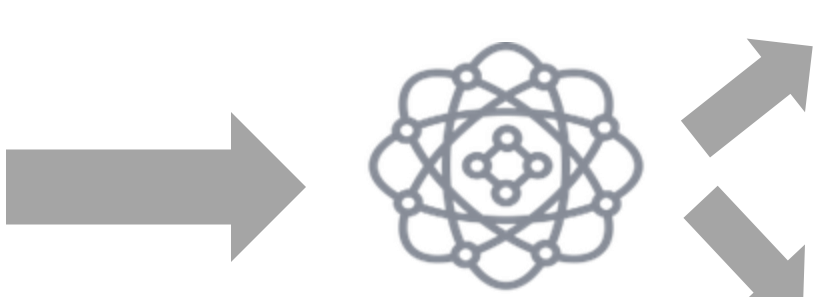
An example of Tabular data

Challenges in applying LLMs to tabular AD(2)

- AD is unsupervised that does not have labels.

Age	Accident Area	Deductible	Vehicle Price
21	Urban	300	More than 69,000
⋮	⋮	⋮	⋮
34	Rural	200	20,000 to 29,000

Normal tabular data



LLM

Normal

Anomalous

Proposed approach: AnoLLM

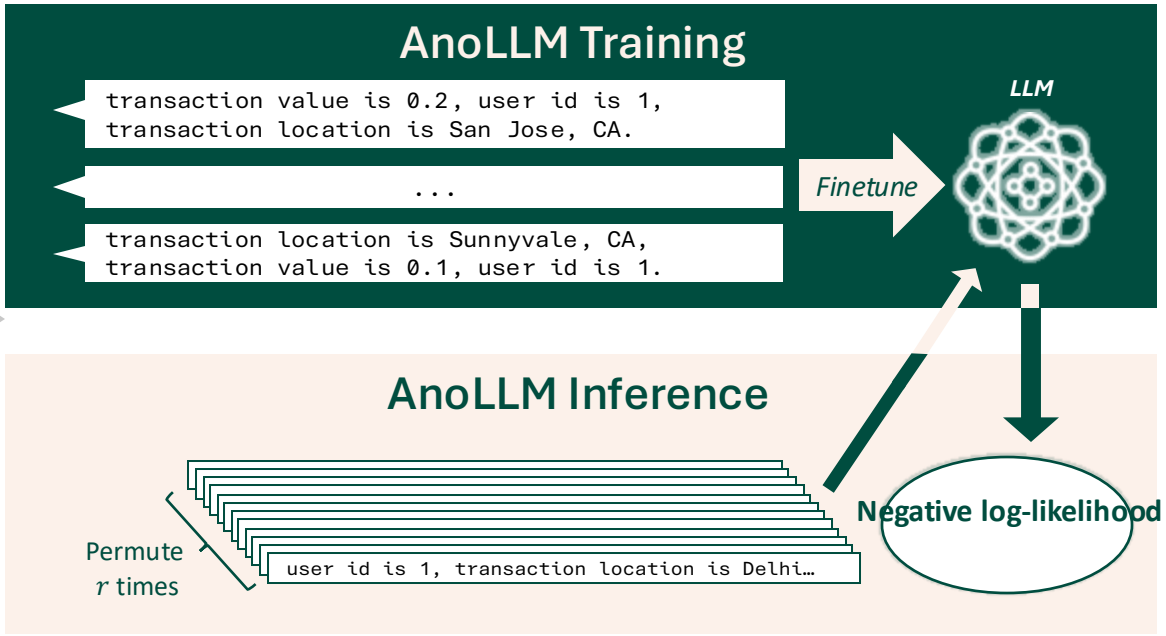
Training Data

User id	Transaction value	Transaction location
1	50.99	San Jose, CA
1	13.00	Palo Alto, CA
1	25.43	Sunnyvale, CA
3	1000.00	Orlando, FL

Test Data

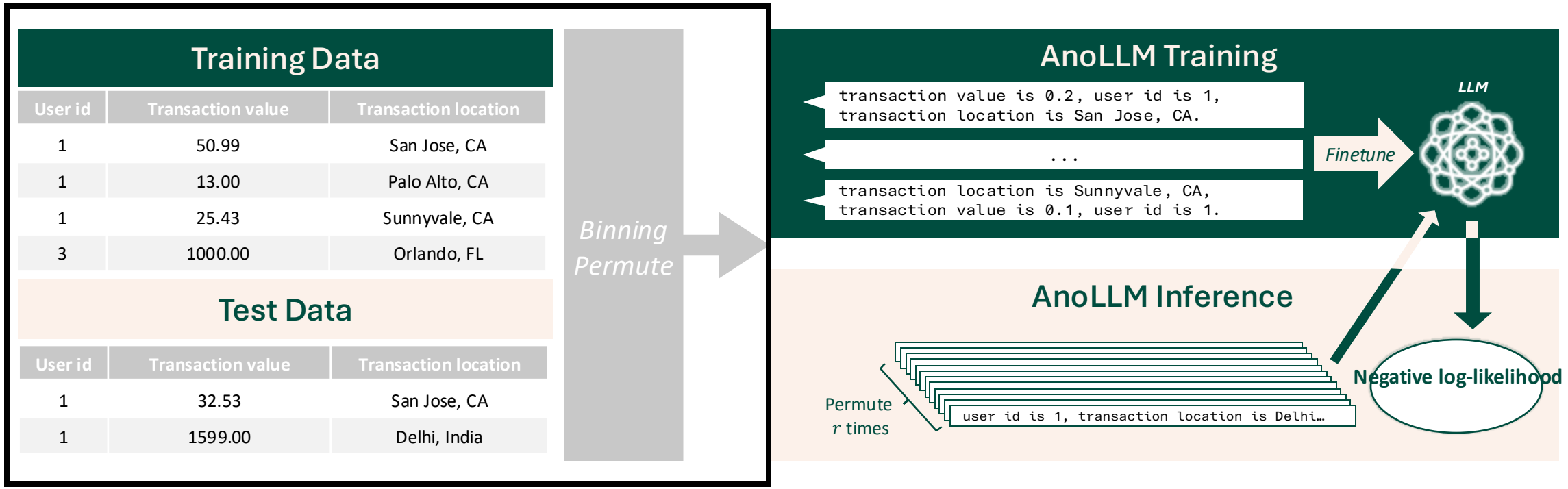
User id	Transaction value	Transaction location
1	32.53	San Jose, CA
1	1599.00	Delhi, India

Binning
Permute



Proposed approach: AnoLLM

Step 1: Serialization



- Transform each row of data into a **sequence** of words.

Proposed approach: AnoLLM

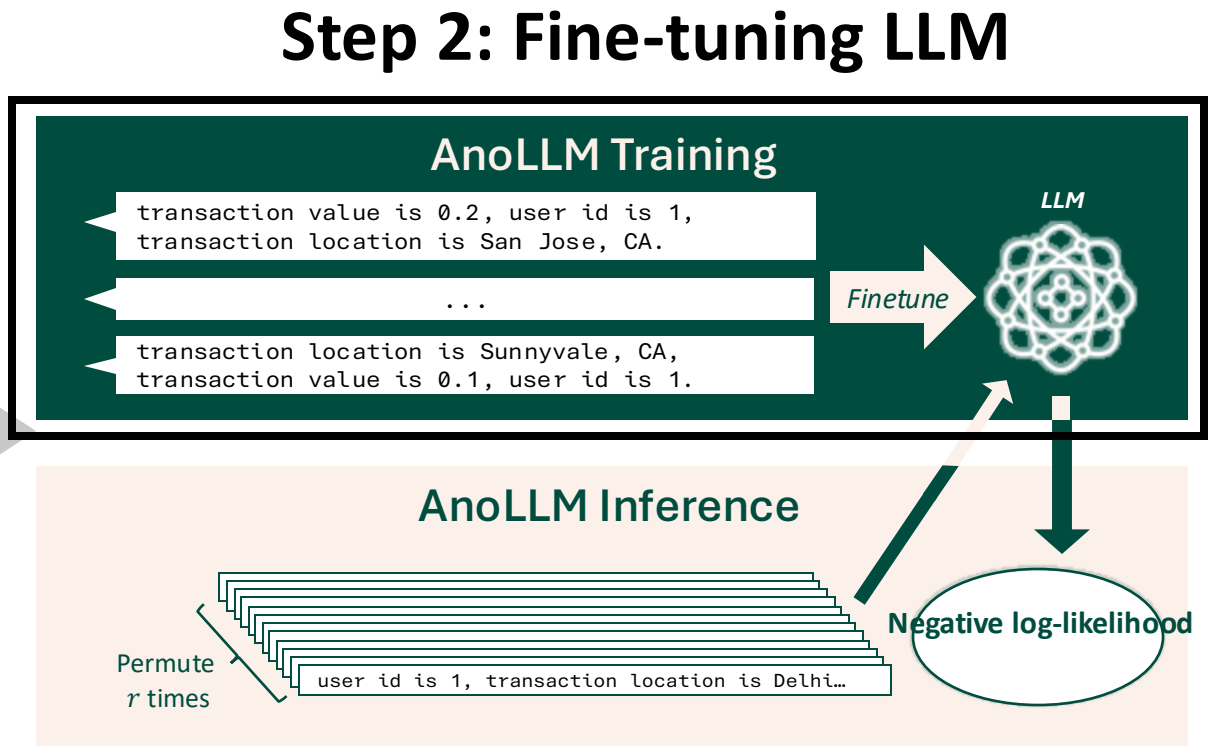
Training Data

User id	Transaction value	Transaction location
1	50.99	San Jose, CA
1	13.00	Palo Alto, CA
1	25.43	Sunnyvale, CA
3	1000.00	Orlando, FL

Test Data

User id	Transaction value	Transaction location
1	32.53	San Jose, CA
1	1599.00	Delhi, India

Binning
Permute



- Fit the serialized tabular data via **next-token-prediction.**

Proposed approach: AnoLLM

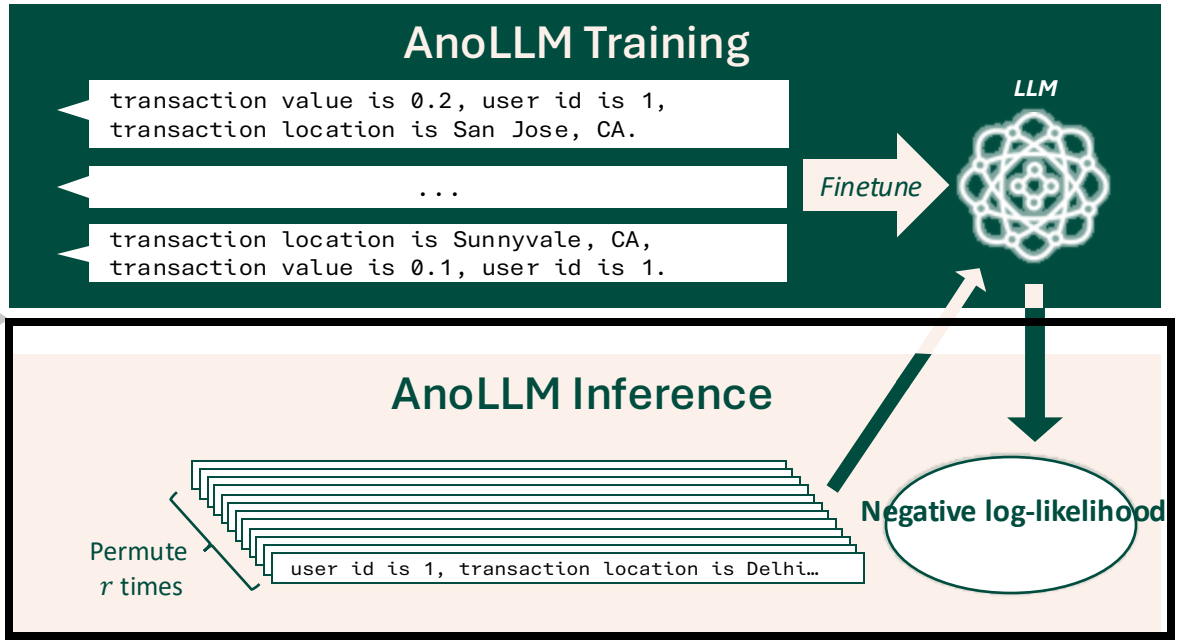
Training Data

User id	Transaction value	Transaction location
1	50.99	San Jose, CA
1	13.00	Palo Alto, CA
1	25.43	Sunnyvale, CA
3	1000.00	Orlando, FL

Test Data

User id	Transaction value	Transaction location
1	32.53	San Jose, CA
1	1599.00	Delhi, India

Binning
Permute



Step 3: Computing anomaly scores

- Use **normalized output probabilities** as anomaly scores.

Experimental Results

Methods \ Datasets	Fake job posts	Fraud ecommerce	Lympho-graphy	Seismic	Vehicle insurance	20news groups	Average
Classical methods							
Iforest	0.755	0.501	0.673	0.692	0.496	0.623	0.623
PCA	0.724	0.647	0.826	0.692	0.509	0.623	0.670
KNN	0.636	**1**	0.860	0.738	0.524	0.605	0.727
ECOD	0.512	0.755	0.830	0.692	0.509	0.62	0.653
Deep learning based methods							
DeepSVDD	0.561	**1**	0.899	0.713	0.505	0.597	0.713
RCA	0.629	**1**	0.919	0.727	0.531	0.546	0.725
SLAD	0.603	0.998	0.964	0.714	0.556	0.64	0.746
GOAD	0.566	0.998	0.817	0.717	0.512	0.63	0.707
NeuTral	0.548	**1**	0.847	0.681	0.507	0.658	0.707
ICL	0.699	**1**	0.827	0.719	0.501	0.671	0.736
DTE	0.548	**1**	0.909	0.714	0.512	0.6	0.714
REPEN	0.653	**1**	0.808	0.724	0.513	0.574	0.712
AnoLLM							
SmolLM-135M	0.800	**1**	0.968	0.712	**0.569**	**0.766**	0.803
SmolLM-360M	**0.814**	**1**	**0.995**	**0.746**	0.555	0.752	**0.810**

Table 2: AUC-ROC scores for all methods on the six datasets containing mixed types of features.

- AnoLLM performs the best on the six datasets containing mixed type of features.

Thanks for listening!

Welcome to our poster for more information!