TransactionGPT

Foundation Model for Payment Transaction Data

Yingtong Dou

Staff Research Scientist @ Visa Research

yidou@visa.com

Yuzhong Chen

Sr. Staff Research Scientist @ Visa Research

yuzchen@visa.com

FOUNDATION MODEL

Revolution of Foundation Models

Foundation models enable broad AI capabilities via large-scale self-supervised training across multiple domains.

Language **Time series** **Tabular data**

…

Complexity of Transaction Data

Payment transaction data is multi-modal, temporal, and tabular (MMTT), containing heterogeneous fields requiring specialized handling.

TransactionGPT (TGPT)

A foundation model of consumer payment data, generating transactions, and supporting downstream tasks such as anomaly detection.

TGPT Architecture Highlights

Modality-Split
3D Architecture

Modality-Fusion
Virtual Token Mechanism

Scalability
Compositional Embedding, Local Attention, and more …

FOUNDATION MODELING TECHNIQUES

Language

Tokens: Text tokens
Method: SFT / RL on pretrained LLM
Example: FinGPT

Event / Behavior

Tokens: Special tokens
Method: Train from Scratch / MLLM
Example: PANTHER

Data / Value

Tokens: Embeddings
Method: Train from Scratch
Example: TransactionGPT

Foundation Models for Predictive Modeling

Data with proprietary domain knowledge and rich priors

Multiple predictive downstream tasks

Strict requirements on robustness, performance, latency, and throughput

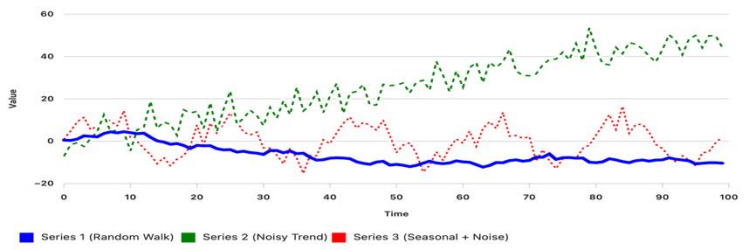
TRANSACTION-LIKE DATA STRUCTURE

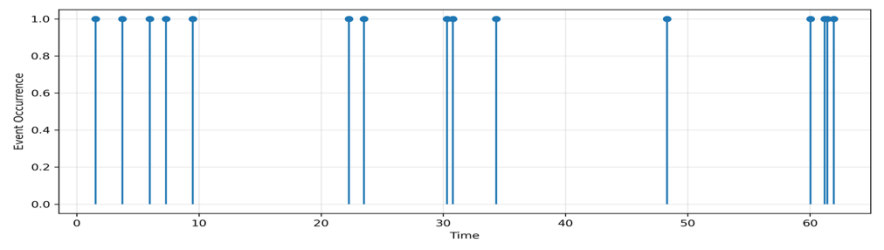
Tabular + Time Series + Point Process
A mixture of data types requiring unified handling

Categorical + Numerical Variables
High cardinality (Merchant ID, Card ID) and low cardinality (MCC, weekday) variables, plus amounts and timestamps

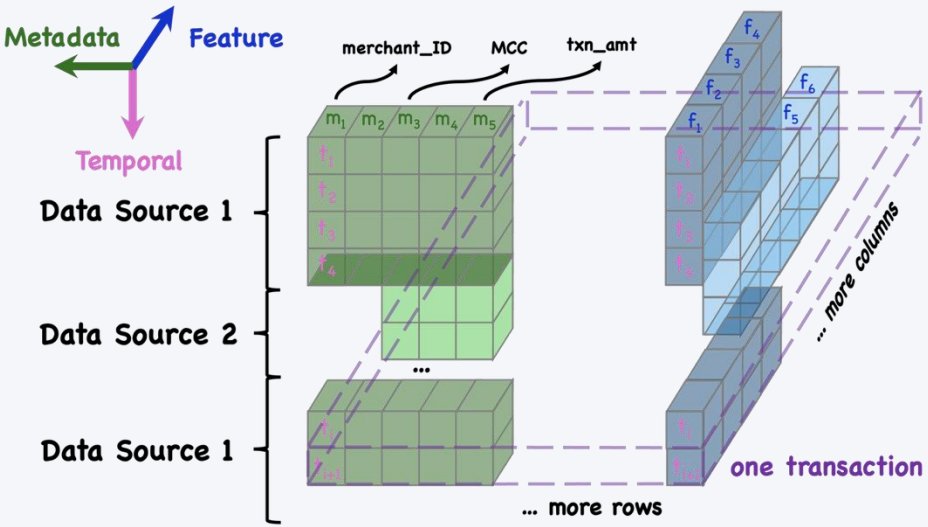
Multiple Data Sources
Different datasets combined for comprehensive transaction modeling

Card	Timestamp	Merchant	MCC	City	State	Amount
12345	12/01/2024	Costco	xxxx	Foster City	CA	$106.00
12345	01/09/2025	Chicken G's	yyyy	Mountain View	CA	$23.99
12345	03/18/2025	Apple	zzzz	Chicago	IL	$900.00
12345	04/30/2025	AMC	ssss	NYC	NY	$21.99





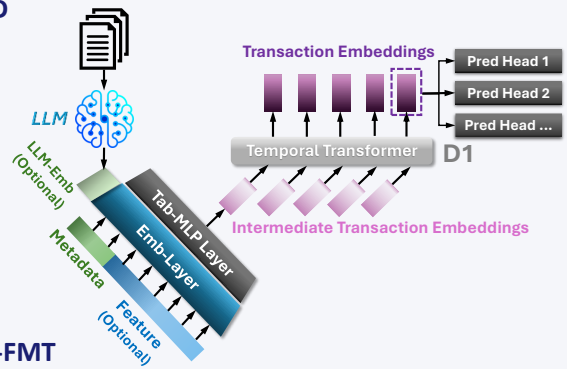
DATA STRUCTURE: MULTI-MODAL-TEMPORAL-TABULAR (MMTT)



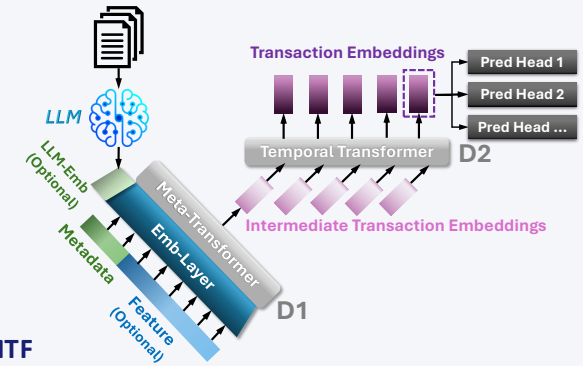
Component	Description
Metadata (M)	Numerical and categorical fields such as amount, timestamp, processing codes
Entities (E)	Merchant ID, merchant category, location information
Features (F)	Task-dependent attributes, often high-dimensional and numerical
Temporal	Sequential ordering of transactions over time

MODEL ARCHITECTURE EVOLUTION

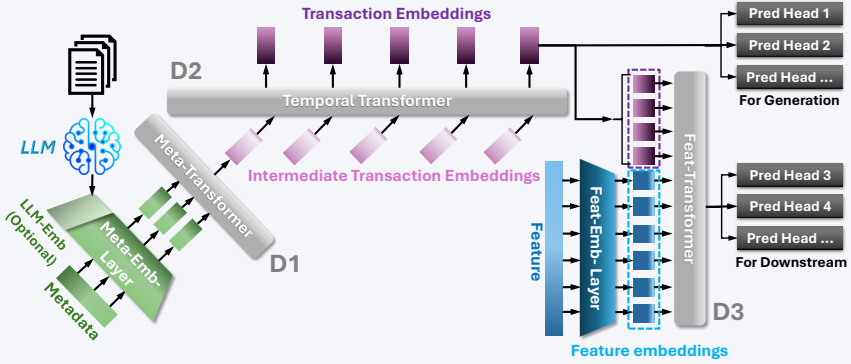
TGPT-1D



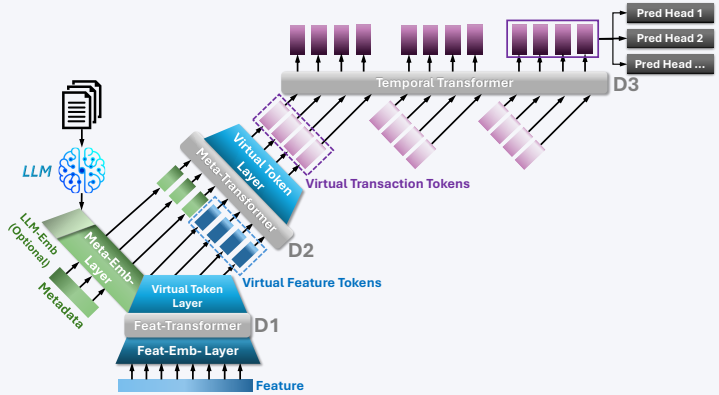
TGPT-2D



TGPT-3D-FMT

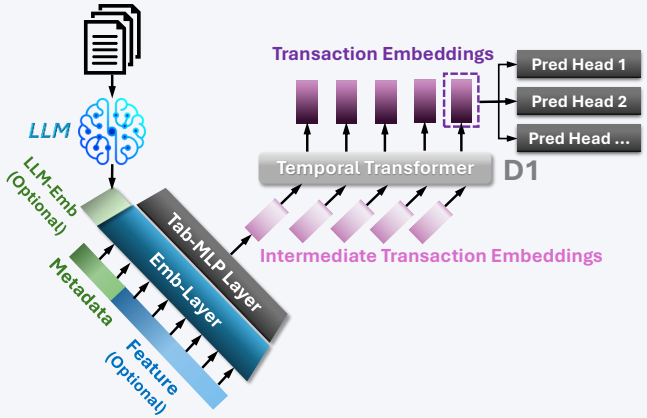


TGPT-3D-MTF

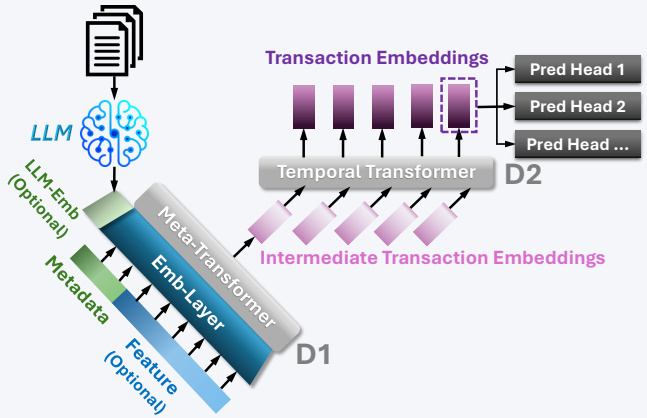


1D & 2D ARCHITECTURE

TransactionGPT-1D



TransactionGPT-2D



Embedding Layer Overview

Data Type	Encoder	Note
Numerical	Linear NN	Log transformation before applying NN
Categorical (low)	Embedding table	Can be pretrained embeddings
Categorical (high)	Compositional embedding	Hashing technique
Time	Linear NN	Multiple fields: relative & absolute time

EXP 1: MCC PREDICTION

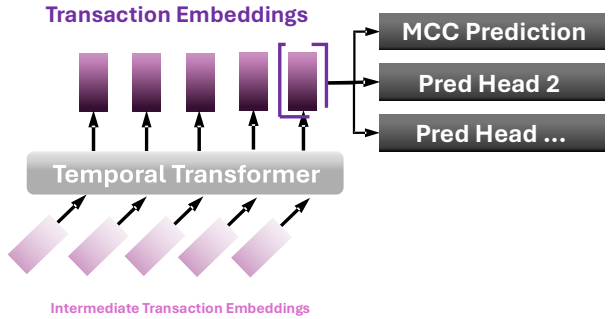
Training/Fine-tuning on 1M seqs, testing on 100K seqs, predicting 800 Merchant Category Code

Same self-supervised information for both LLM and TGPT

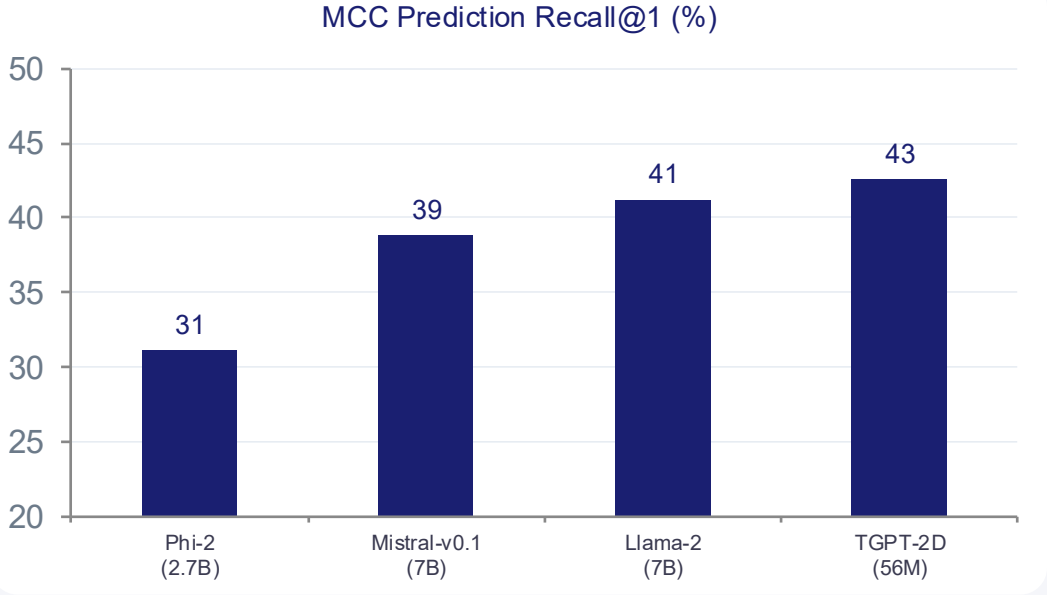
LLM-based Prediction

[More historical transaction info...]
Target field: **{field_value}**, transaction amount: **{amount}**, transaction date: **{date}**, ⋯,
transaction time: **{time}**. ▷ last transaction info.
Target field: [MASK]. ▷ field to be predicted.

TransactionGPT Prediction



MCC PREDICTION RESULTS



Efficiency Advantage

92%

fewer parameters
(56M vs 7B)

300×

faster inference
(0.27ms vs 84.9ms)

On a single NVIDIA A100 GPU (80GB)

EXP 2: RECOMMENDATION & TRAJECTORY PREDICTION

Training on 200M len=16 seqs, testing on 20M seqs, predicting 500K unique restaurant merchants..

Restaurant Prediction Performance

Metric	SASRec	TGPT-1D	TGPT-2D
Rec@1	11.9%	12.8%	**14.2%**
Rec@50	38.5%	42.5%	**45.6%**

TGPT-2D ranks the exact future restaurant in the top 50 candidates out of 500K in 45.6% of test cases.

Location Prediction Accuracy

84% State (Top-1)

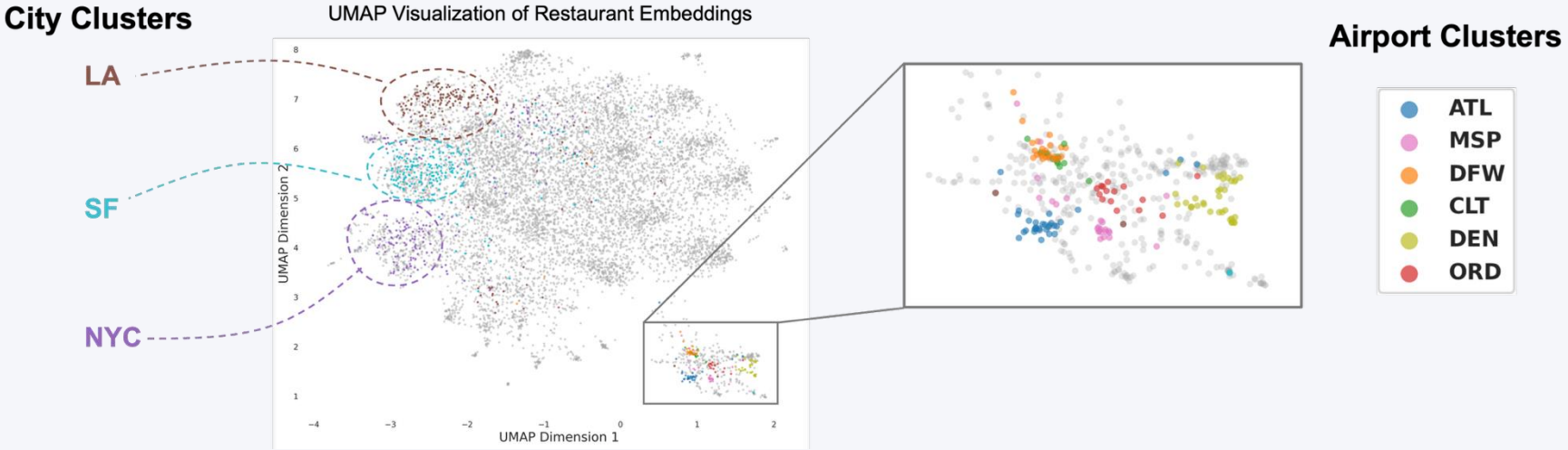
69% City (Top-10)

28% Zip (Top-1)

TGPT predicts future dining locations without ingesting any location data — purely from transaction patterns.

EXP 3: RESTAURANT EMBEDDING VISUALIZATION

UMAP visualization of TGPT merchant embeddings of top 10K restaurants ranked by transaction volume



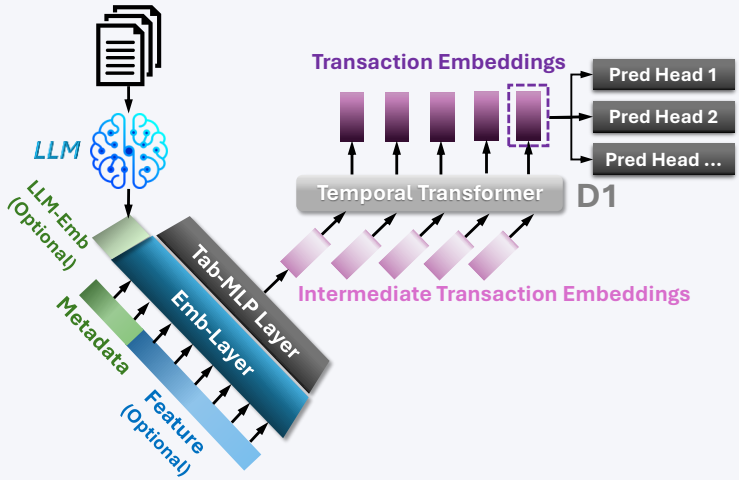
ANOMALY TRANSACTION DETECTION

Binary classification of a single transaction with downstream features available

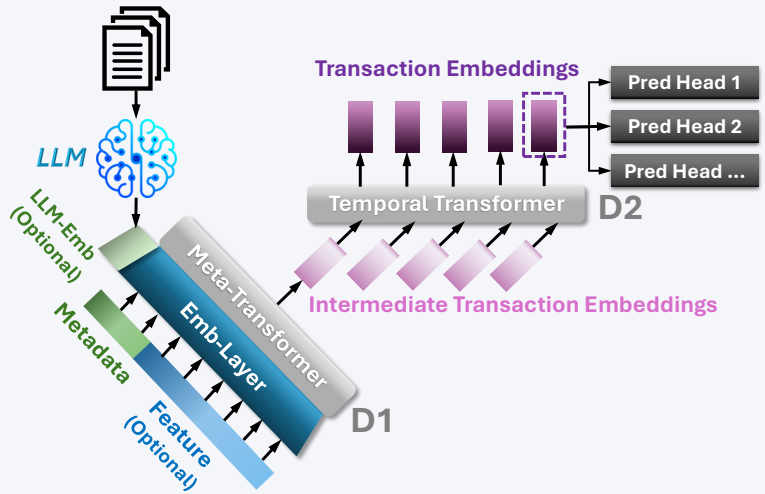
Model	AD (%)	MCC (%)	Mrch (%)	Amount		Time	
	RI↑	Rec@1↑	Rec@1↑	MAE↓	MSE↓	MAE↓	MSE↓
Transformers4Rec	-	33.46	-	-	-	-	-
Feat-Transformer	+7.7	-	-	-	-	-	-
TGPT-1D							
- w/o Feat	-90.5	48.86	31.60	1.29	3.84	0.1350	0.0355
- w/ Feat	-9.1	49.25	30.46	1.30	3.93	0.1370	0.0282
TGPT-2D							
- w/o Feat	-87.0	48.89	31.41	1.27	3.78	0.0702	0.0100
- w/ Feat	+7.3	49.17	30.38	1.26	3.79	0.0740	0.0095

BOTTLENECK OF 1D & 2D

TransactionGPT-1D



TransactionGPT-2D



Key Bottlenecks

Embedding size conflict

Metadata/Entities need large embedding size but have small amount.

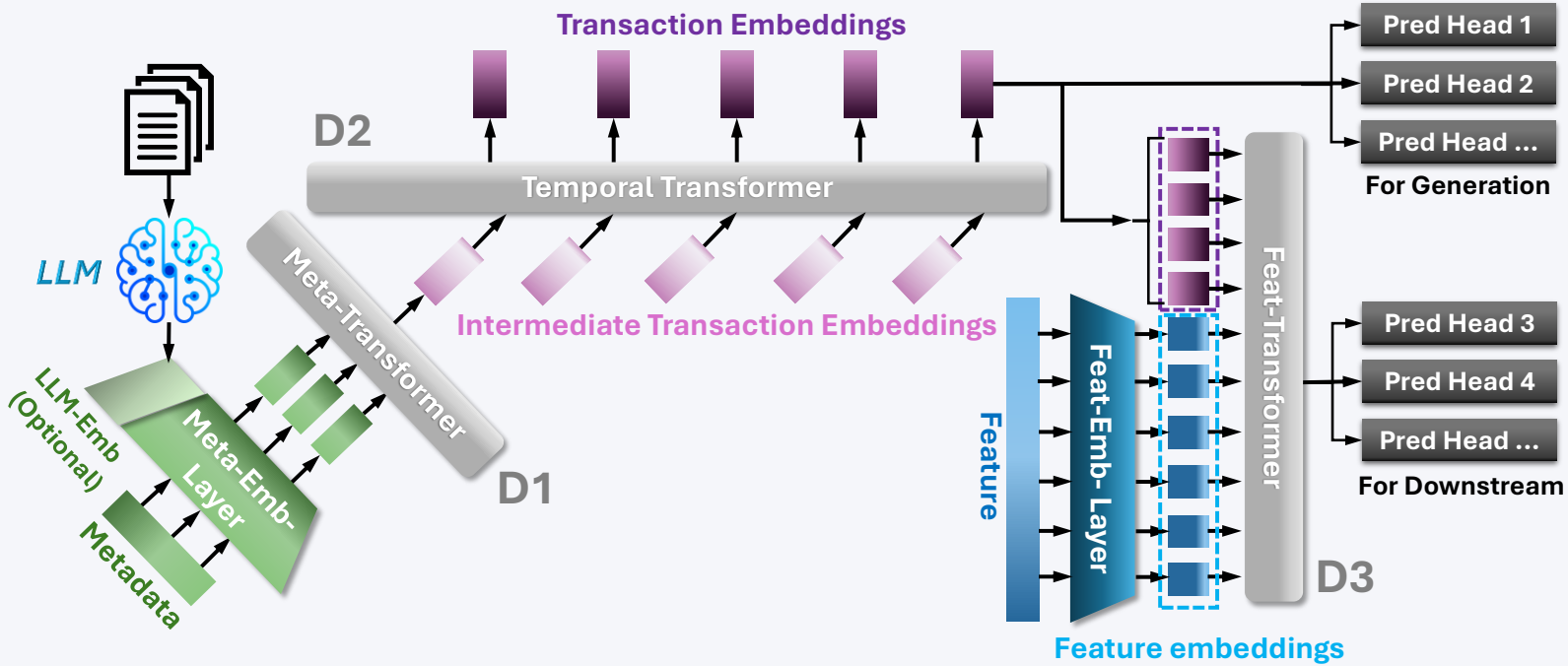
Features need small embedding size but have large amount.

Transaction level integration issues

The explosion of number of tokens when combining all modalities at the transaction level.

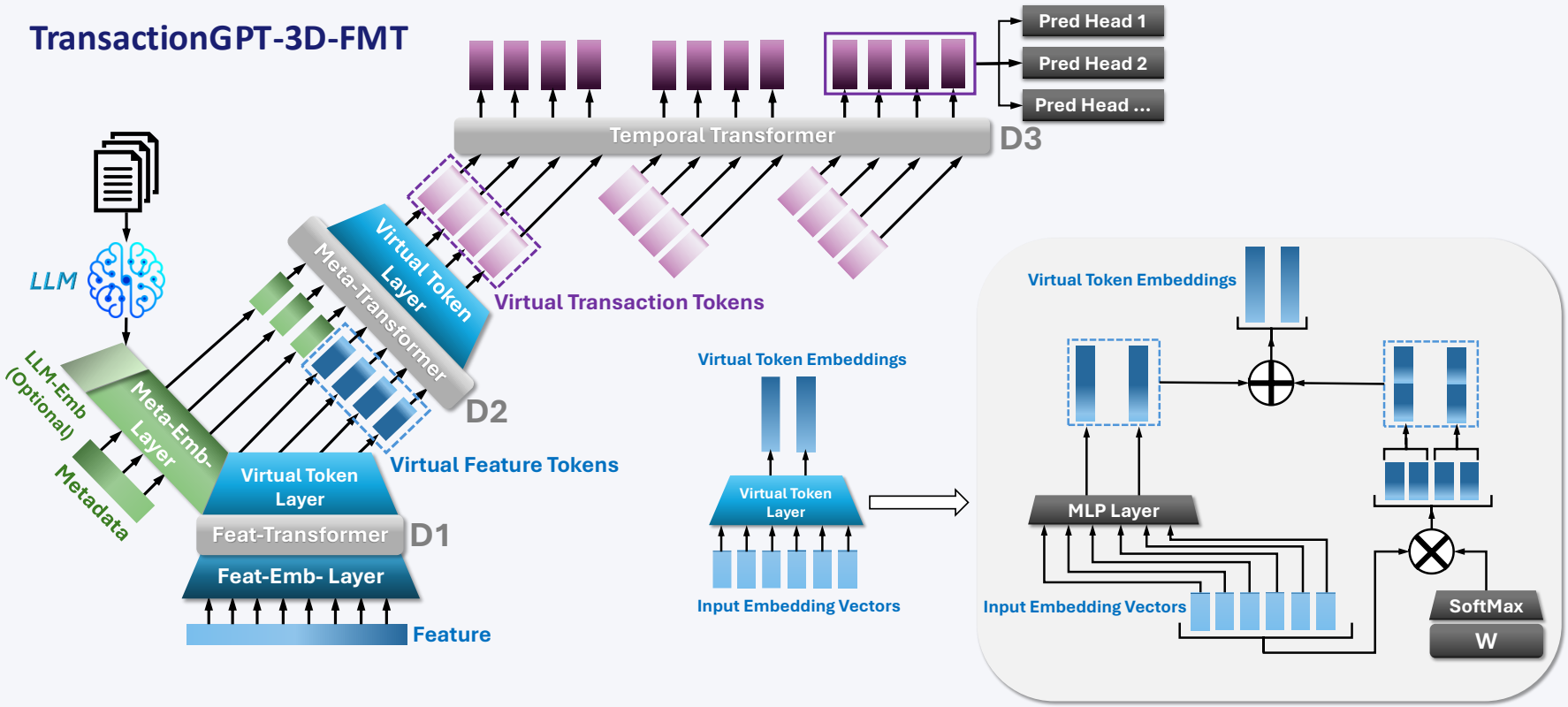
MODALITY SPLIT: 3D ARCHITECTURE

TransactionGPT-3D-MTF



3D-FMT WITH VIRTUAL TOKEN MECHANISM

TransactionGPT-3D-FMT



ANOMALY DETECTION: PERFORMANCE COMPARISON

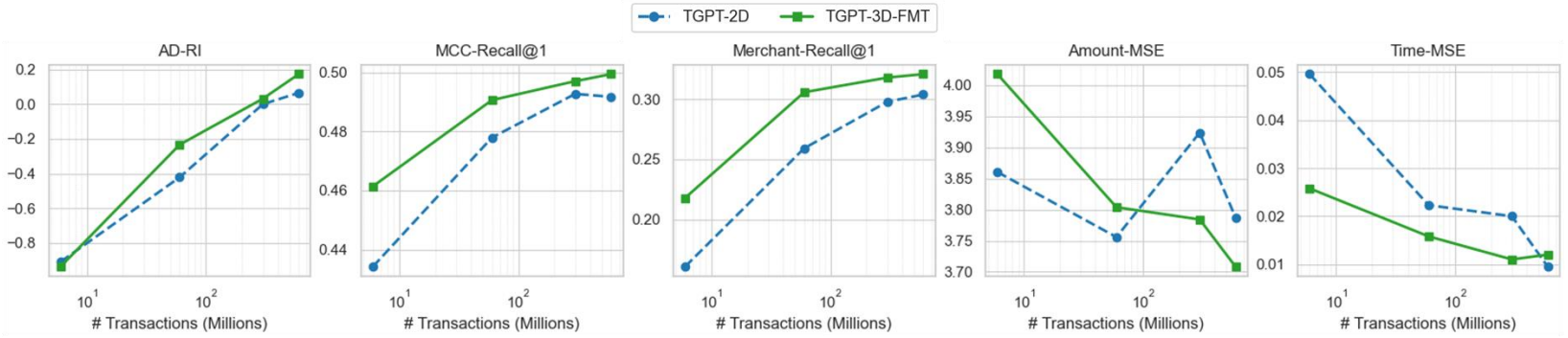
Performance comparison of TGPT and its variants on transaction generation and classification tasks

Table 4: Performance comparison of TGPT and its variants on transaction generation and anomaly detection (**AD**) tasks. For compliance, only the relative improvement (**RI**) over the production model performance is shown for the anomaly detection task.

Model	**AD (%)**	**MCC (%)**	**Mrch (%)**	**Amount**		**Time**	
	RI↑	**Rec@1↑**	**Rec@1↑**	**MAE↓**	**MSE↓**	**MAE↓**	**MSE↓**
Transformers4Rec	-	33.46	-	-	-	-	-
Feat-Transformer	+7.7	-	-	-	-	-	-
TGPT-1D							
- w/o Feat	-90.5	48.86	31.60	1.29	3.84	0.1350	0.0355
- w/ Feat	-9.1	49.25	30.46	1.30	3.93	0.1370	0.0282
TGPT-2D							
- w/o Feat	-87.0	48.89	31.41	1.27	3.78	0.0702	0.0100
- w/ Feat	+7.3	49.17	30.38	1.26	3.79	0.0740	0.0095
TGPT-3D-MTF							
- Segments	+14.6	48.23	32.41	1.33	4.05	0.1420	0.0329
- MLP-Scaling	+9.5	48.17	32.29	1.32	4.05	0.1400	0.0325
TGPT-3D-FMT							
- FMVTL	+19.2	49.94	32.08	**1.23**	**3.71**	0.0893	0.0120
- FMVTL-2	+15.5	**50.12**	32.70	1.24	3.73	0.0800	0.0098
- FMVTL-nonlin	+6.7	49.55	31.20	1.30	3.86	0.0901	0.0141
- FMVTL-lin-map	+0.2	49.62	31.06	1.26	3.78	0.0880	0.0152
- FVTL	-12.9	48.45	30.91	1.28	4.16	0.0767	0.0111
- FMVTL + LLM	**+22.5**	50.01	**32.73**	1.25	3.77	**0.0686**	**0.0072**

DATA SCALING of TGPT

Performance metrics as a function of the number of transactions trained on



KEY TAKEAWAYS

1

Impact of Foundation Models
Information across different domains and modalities jointly benefit the downstream tasks

2

Importance of Information Fusion
Maintain the optimal information bandwidth across dimensions and modalities

3

Extra Signal from LLMs
MCC embeddings generated by LLMs help the anomaly transaction detection

4

Design Choices for Transformers
Please check our technical report for more details

Thank You!

https://arxiv.org/pdf/2511.08939

Core Contributors

Yingtong Dou, Zhimeng Jiang, Tianyi Zhang, Mingzhi Hu, Zhichao Xu, Yuzhong Chen

Contributors

Shubham Jain, Uday Singh Saini, Xiran Fan, Jiarui Sun, Menghai Pan, Junpeng Wang, Xin Dai, Liang Wang, Chin-Chia Michael Yeh, Yujie Fan, Yan Zheng, Vineeth Rakesh, Huiyuan Chen, Guanchu Wang, Mangesh Bendre, Zhongfang Zhuang, Xiaoting Li, Prince Aboagye, Vivian Lai, Minghua Xu, Hao Yang, Yiwei Cai, Mahashweta Das

Project Lead

Yuzhong Chen