LLM4ES: Learning User Embeddings from Event Sequences via Large Language Models

Aleksei Shestov[*] Omar Zoloev[*] Maksim Makarenko[*] Mikhail Orlov[*] Egor Fadeev
Ivan Kireev[*] Andrey Savchenko[*†]

[*]Sber AI Lab [†]HSE University

CIKM 2025

Motivation

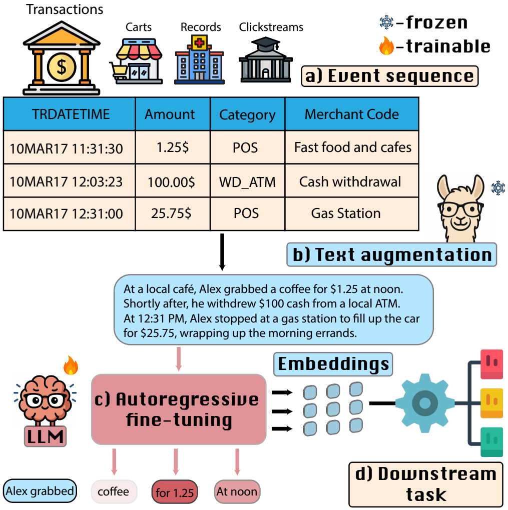
- Event sequence (ES) data is ubiquitous: finance, healthcare, e-commerce, education
- Challenge: How to effectively construct user embeddings from ES data?
- Traditional approaches:
  - Contrastive learning (CoLES)
  - Next event prediction (NPPR)
  - Temporal point processes (DeTPP, IFTPP)
- **Key insight**: LLMs encode rich knowledge about event semantics and behavior patterns
- **Challenge**: Adapting LLMs to structured event sequence data

Our Contribution

Main Contributions

1. **LLM4ES Framework**: Novel approach for extracting user embeddings by fine-tuning LLMs on domain-specific event sequences
   - Up to 7% relative improvement in ROC-AUC over SOTA methods
   - Works both standalone and in ensemble
2. **Text Enrichment Technique**: General method for adapting LLMs to low-variability event sequence data
   - Up to 1.8% relative improvement over raw serialization

LLM4ES Framework Overview



(a) Event sequences → Text serialization

(b) Text enrichment via LLaMA-3.1-8B-instruct

(c) Fine-tune LLaMA-3.2-3B with next-token prediction

(d) Extract embeddings for downstream tasks

Method: Event Sequence Serialization

Step 1: Convert structured events to text format

Serialization Format

- Header line: feature names separated by "—"
- Each line: one event with comma-separated values
- Replace ordinal codes with textual descriptions
  - Example: MCC=5411 → "Grocery Stores, Supermarkets"

Example

```
scaled time | amount(rubles) | trx date | Merchant category
7, 3.79, 728, Maintenance of existing car (gas station)
7, 4.25, 728, Chain supermarkets and grocery stores
```

Method: Text Enrichment

Step 2: Enrich serialized text for better LLM adaptation

Why Text Enrichment?

- Problem: Low variability in serialized event sequences
- Solution: Use LLaMA-3.1-8B-instruct to rewrite events into diverse formats
- Formats: JSON, Markdown, HTML, XML, YAML, tables, narratives

Enrichment Result

Transaction Breakdown

Grocery Stores: 24 transactions, 166 rubles

Service Stations: 7 transactions, 33 rubles

Date	Amount	Category
17298	4 rubles	Grocery Stores

Method: LLM Fine-tuning & Embedding Extraction

Step 3: Fine-tune LLM on enriched text

Next-Token Prediction Loss

$$\mathcal{L}_{\text{NTP}} = -\sum_{t} \log P(x_t \mid \mathbf{x}_{<t}; \theta)$$

- Full fine-tuning (not LORA) of LLaMA-3.2-3B

Step 4: Extract embeddings

Mean Pooling Over Last Layers

$$\textbf{User}_{emb} = \text{MeanPooling}\left(\frac{1}{k}\sum_{l=L-k}^{L-1} \mathbf{H}_l\right)$$

- Averaging $k = 8$ last layers, mean pooling over tokens

Experimental Setup

Datasets

- **Financial**: Rosbank (churn), Gender, Age
- **Non-financial**: MovieLens 100k, MIMIC-IV
- **Internal**: Acquiring with CLTV target

Validation Strategy

- 10% test set, 90% train
- 5-fold prediction
- LightGBM classifier on embeddings
- Metric: ROC-AUC

Baselines

- **ES**: CoLES, NPPR
- **TPP**: DeTPP, IFTPP
- **NLP**: NSP, RTD
- **Tabular**: Aggregated features

Dataset	Labeled	Features
Rosbank	5,217	5
Gender	5,000	5
Age	30,000	4

Main Results: Financial Datasets

Table: ROC-AUC scores on financial datasets

Method	Rosbank	Gender	Age
Agg	0.827 ± 0.010	<u>0.877</u> ± 0.004	0.629 ± 0.002
CPC	0.792 ± 0.015	0.851 ± 0.006	0.602 ± 0.004
RTD	0.771 ± 0.016	0.855 ± 0.008	0.631 ± 0.006
CoLES	*0.841* ± 0.005	**0.882** ± 0.004	<u>0.644</u> ± 0.005
NSP	0.828 ± 0.012	0.852 ± 0.011	0.621 ± 0.005
NPPR	<u>0.845</u> ± 0.003	-	*0.642* ± 0.001
DeTPP	0.823 ± 0.002	0.850 ± 0.005	0.632 ± 0.004
IFTPP	0.828 ± 0.004	0.863 ± 0.003	0.632 ± 0.003
IFTPP-T	0.814 ± 0.004	0.852 ± 0.005	0.620 ± 0.002
LLM4ES	**0.851** ± 0.004	*0.875* ± 0.004	**0.659** ± 0.004

- **Best performance** on Rosbank (+0.7% over NPPR) and Age (+2.3% over CoLES)
- **Age dataset**: +10% relative gain over multiple baselines

Ensemble Performance

Method	Gender	Age
CoLES	0.882	0.644
CoLES + Agg	0.890	0.629
CoLES + IFTPP	0.884	0.638
CoLES + IFTPP-T	0.880	0.635
CoLES + LLM4ES	**0.892**	**0.665**

Table: Ensemble results (concatenated embeddings)

Key Findings

- CoLES + LLM4ES: **+1.1%** on Gender, **+3.3%** on Age vs. standalone CoLES
- **+4.2%** over CoLES + IFTPP on Age
- LLM4ES captures **complementary information**

Cross-Domain Results

Table: Performance beyond financial domains and internal dataset

Model	MovieLens 100k Gender	MovieLens 100k Age (MAE)	MIMIC-IV ROC-AUC	CLTV MAE
CPC	0.567	9.97	0.8855	1.62×10^4
CoLES	0.683	8.70	0.9094	1.6×10^4
LLM4ES	**0.748**	**7.34**	**0.9536**	$\mathbf{1.56 \times 10^4}$

- **MovieLens**: +9.5% gender classification, -15.6% age MAE
- **MIMIC-IV**: +4.9% mortality prediction
- **CLTV**: Improved MAE on proprietary dataset

Ablation Study: Component Analysis

Table: Impact of LLM4ES components

LLM	Fine-tuning Data	Rosbank	Gender
Pre-trained	-	0.823	0.711
Fine-tuned	Original serialized	0.839	0.850
Fine-tuned	Text enriched	*0.849*	*0.867*
Fine-tuned	Text enriched (both datasets)	**0.851**	**0.875**

- Fine-tuning on serialized data: +1.7% (Rosbank), +13.9% (Gender)
- Text enrichment adds: +1.0% (Rosbank), +1.7% (Gender)
- Cross-dataset training further improves results

Ablation Study: Text Enrichment Analysis

Table: Text enrichment strategies (Rosbank dataset)

Method	Data Amount	ROC-AUC
Raw serialization (TALLRec)	1x	0.839 ± 0.003
Best single format (HTML)	1x	0.845 ± 0.002
Worst single format (JSON)	1x	0.839 ± 0.006
Mixed formats	1x	**0.849** $\pm$ 0.001
Mixed formats	7x	*0.854* $\pm$ 0.004
Mixed + raw	8x	0.843 ± 0.003
LLM-TRSR	1x	0.664 ± 0.014
HKFR	1x	0.756 ± 0.008
LLM-TRSR, only train	1x	0.847 ± 0.005
HKFR, only train	1x	0.823 ± 0.006

- Mixed-format enrichment outperforms single formats
- Scaling enrichment volume improves performance
- Outperforms prior LLM-based approaches (HKFR, LLM-TRSR, TALLRec)

LLM Architecture Ablation

Table: Text enrichment on different LLMs (Rosbank dataset)

Base LLM	Data representation	ROC-AUC
Qwen 3 1.7b	Original Serialized	0.834
Qwen 3 1.7b	Text-enriched	<u>0.844</u>
LLaMA 3.2 3b	Original Serialized	0.839 ± 0.003
LLaMA 3.2 3b	Text-enriched	$\mathbf{0.849} \pm \mathbf{0.001}$

- Text enrichment: $+1\%$ for both Qwen and LLaMA models.

Data Efficiency Analysis

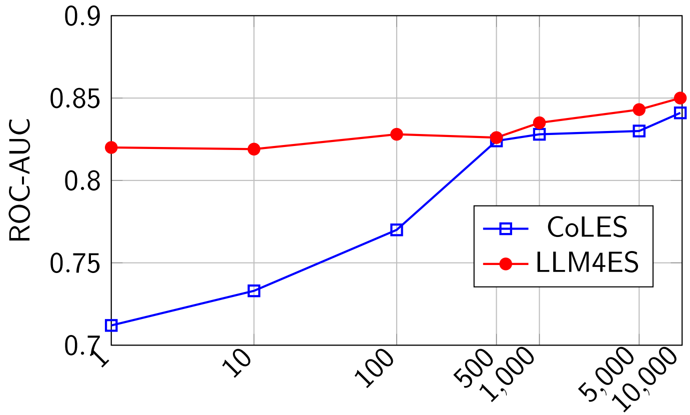


Figure: Performance vs. sample size on Rosbank dataset

LLM4ES excels in low-data regimes:

- +15.2% at 1 sample
- +8.6% at 10 samples

Conclusion & Future Work

Summary

LLM4ES: Novel framework for learning user embeddings from event sequences via LLMs

- Leverages pre-trained LLM knowledge about event semantics
- Text enrichment bridges gap between LLM training and structured data
- Achieves SOTA results both standalone and in ensemble

Future Directions

- More context-effectient representation
- More sophisticated embedding extraction methods
- Investigate other enrichment strategies
- Scale to larger LLMs and datasets

Code available at: `https://github.com/tsebaka/LLM4ES`

Thank You!

Questions?

Contact:
shestovmsu@gmail.com
ozoloevwork@gmail.com