Enhancing Foundation Models in Transaction Understanding with LLM-based Sentence Embeddings

Xiran Fan Zhimeng Jiang Chin-Chia Michael Yeh
Yuzhong Chen Yingtong Dou Menghai Pan Yan Zheng
Visa Research
Foster City, CA
{xirafan, zhimjian, miyeh, yuzchen, yidou, menpan, yazheng}@visa.com

Abstract

The ubiquity of payment networks generates vast transactional data encoding rich consumer and merchant behavioral patterns. Recent foundation models for transaction analysis process tabular data sequentially but rely on index-based representations for categorical merchant fields, causing substantial semantic information loss by converting rich textual data into discrete tokens. While Large Language Models (LLMs) can address this limitation through superior semantic understanding, their computational overhead challenges real-time financial deployment. We introduce a hybrid framework that uses LLM-generated embeddings as semantic initializations for lightweight transaction models, balancing interpretability with operational efficiency. Our approach employs multi-source data fusion to enrich merchant categorical fields and a one-word constraint principle for consistent embedding generation across LLM architectures. We systematically address data quality through noise filtering and context-aware enrichment. Experiments on large-scale transaction datasets demonstrate significant performance improvements across multiple transaction understanding tasks.

1 Introduction

Foundation models have achieved remarkable success across diverse domains, from natural language processing Brown et al. [2020], Devlin et al. [2018] and computer vision Dosovitskiy et al. [2020], Awais et al. [2025] to multimodal learning Ramesh et al. [2021], Alayrac et al. [2022] and recommendation systems Huang et al. [2024]. Despite this progress, the development of foundation models for tabular data ubiquitous in real-world applications (e.g., transaction understanding) remains comparatively underexplored. Advancement is hindered by intrinsic challenges such as permutation invariance, heterogeneous feature types, and domain-specific semantics. Recent work by Yeh et al. [2025] takes a step in this direction by modeling transactional data via sequential patterns across transactions. However, it encodes several categorical attributes as discrete indices, discarding rich semantics; for example, mapping merchant name "Costco" to an index elides its wholesale, membership-based identity. We argue that large language models (LLMs), with broad world knowledge, offer a promising path forward: they can transform textual fields (e.g., merchant names and locations) into semantically meaningful embeddings while respecting production constraints on latency, compute budgets, and robustness to real-world noise.

In this paper, we propose a practical framework that uses LLM-generated sentence embeddings to initialize categorical field representations in a foundation model for tabular data, mitigating information loss while remaining production-viable. We focus on fields with enrichable context in real-world transaction records—merchant category codes (MCC), merchant names, and locations—and we fuse multiple sources to construct prompts optimized for consistent embedding generation. A key

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: NeurIPS 2025 Workshop on Generative AI in Finance.

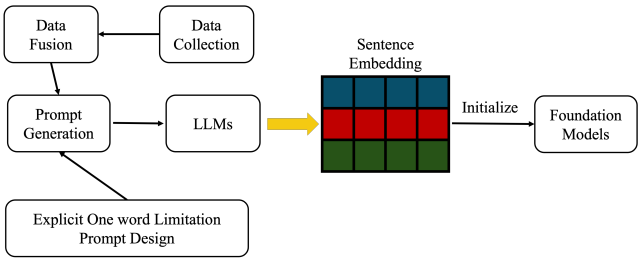

Figure 1: An overview of proposed method.

design is the one-word limitation principle Jiang et al. [2024b]; for decoder-only models, prompts like "This sentence: '[text]' means in one word:" yield focused outputs. We extract sentence embeddings from the final layer's hidden states (last non-padding token). Our pipeline further applies rule-based filtering, strategic null-token replacement, and retrieval to augment field context. Initialized with these LLM-based embeddings, production models inherit semantic understanding and retain deployment efficiency via task-specific fine-tuning, yielding significant improvements in transaction metrics across multiple LLM architectures.[1]

Our contributions are: (1) an LLM-embedding integration that reduces index-based information loss; (2) a preprocessing and prompt pipeline featuring the one-word limitation principle; (3) multi-source data fusion for categorical enrichment; (4) empirical gains on real-world transaction understanding.

2 Related Work

Foundation models for tabular data Yang et al. [2023], Van Breugel and Van Der Schaar [2024], Zhang et al. [2023] and TabPFN Hollmann et al. [2022] extend pretraining to tables, but mainly address static rows rather than temporal purchasing dynamics. Sequential transaction modeling Skalski et al. [2023] highlights temporal dependencies and cross-transaction relations. ID-based embeddings offer collaborative signals Schafer et al. [2007] but lack semantics and face cold-start issues. Hybrid CF–semantic approaches (e.g., He et al. [2025]) demonstrate the value of textual priors. LLMs have been adapted to structured data (e.g., Su et al. [2024]), but full LLM inference is typically too costly for real-time payments. Our work differs by precomputing LLM sentence embeddings solely for initialization, preserving efficiency while improving semantic coherence.

3 Methodology

Our framework addresses the semantic information loss in traditional foundation models for transaction analysis through a systematic pipeline that leverages LLM capabilities while maintaining production deployment feasibility. Figure 1 illustrates the end-to-end architecture, which consists of interconnected components designed to transform sparse categorical merchant data into semantically rich embeddings. The pipeline has five stages:

- Data collection: merchant name, MCC, and location fields.
- Multi-source enrichment: official MCC descriptions, internal merchant metadata, and location context reduce sparsity and noise.
- Prompt generation: compact, field-specific prompts with an explicit one-word limitation stabilize outputs across LLMs.
- LLM processing: sentence embeddings are extracted from the final hidden layer (last non-padding token) for each field instance.
- Semantic initialization: the extracted vectors initialize the model's categorical embeddings; subsequent end-to-end fine-tuning adapts to tasks.

[1]Transaction volume, decline rate, and fraud rate are commonly used metrics for analyzing transactions in payment network companies.

Table 1: Performance comparison of various LLM architectures (**Model**) and embedding initialization methods (**Emb.**) on transaction prediction tasks. "Vanilla" denotes the original foundation model. Grey-highlighted cells indicate settings where the model outperforms the vanilla baseline.

Model	Emb.	Next Amount		Next MCC		Next City		Next Merchant	
		MAE	sMAPE	Acc	F1	Acc	F1	Acc	F1
Vanilla		37.0430	0.3952	0.4107	0.1118	0.8454	0.6721	0.0760	0.0037
Llama2-7b	MCC	36.7474	0.3954	0.4134	0.1143	0.8443	0.6683	0.0913	0.0099
Llama2-13b		37.2129	0.3978	0.4118	0.1148	0.8436	0.6652	0.0905	0.0103
Llama3-8b		37.5144	0.4010	0.4132	0.1131	0.8456	0.6803	0.0960	0.0103
Mistral-7b		37.0173	0.3961	0.4152	0.1108	0.8449	0.6623	0.0934	0.0100
Llama2-7b	Merchant	36.9498	0.3949	0.4119	0.1054	0.8442	0.6625	0.0938	0.0098
Llama2-13b		37.0119	0.3962	0.4123	0.1014	0.8437	0.6597	0.0898	0.0085
Llama3-8b		37.0511	0.3968	0.4115	0.1102	0.8434	0.6606	0.0991	0.0098
Mistral-7b		36.8613	0.3957	0.4148	0.1141	0.8456	0.6649	0.0930	0.0095
Llama2-7b	MCC Merchant	37.0336	0.3986	0.4145	0.1177	0.8452	0.6603	0.0929	0.0103
Llama2-13b		36.8841	0.3932	0.4154	0.1162	0.8451	0.6536	0.0946	0.0100
Llama3-8b		36.8522	0.3957	0.4168	0.1202	0.8458	0.6636	0.0994	0.0110
Mistral-7b		37.1708	0.3983	0.4126	0.1052	0.8433	0.6462	0.0901	0.0083
Llama2-7b	State City	37.3200	0.4025	0.4163	0.1157	0.8450	0.6642	0.0959	0.0101
Llama2-13b		36.7465	0.3935	0.4151	0.1124	0.8459	0.6646	0.0951	0.0120
Llama3-8b		36.7652	0.3952	0.4146	0.1187	0.8451	0.6607	0.0947	0.0106
Mistral-7b		36.9257	0.3972	0.4125	0.1076	0.8437	0.6577	0.0904	0.0090
Llama2-7b	All Fields	37.0218	0.3977	0.4140	0.1178	0.8462	0.6683	0.0942	0.0098
Llama2-13b		37.1253	0.4003	0.4152	0.1184	0.8454	0.6651	0.0991	0.0115
Llama3-8b		36.8128	0.3927	0.4155	0.1208	0.8455	0.6716	0.0979	0.0110
Mistral-7b		36.8761	0.3955	0.4108	0.0960	0.8434	0.6398	0.0979	0.0074

Data Fusion and Prompting. We integrate official MCC descriptions ISO [2023] with internal merchant and location records, retaining only signals that improve semantic coverage with minimal overhead (e.g., short category text, nearby categories, coarse regional info). Prompts are field-specific and enforce a one-word (or single-token) answer when feasible, reducing verbosity and improving embedding comparability. As shown in Listing 3, we provide a prompt template for generating MCC embeddings. Full templates and examples are provided in Appendix A.

```
Input: MCC "5044"
Prompt: "The MCC 5044, titled 'Photographic, Photocopy, Microfilm Equipment and Supplies', serves
    ↪ business-to-business distributors of office and photographic equipment including film, cash
    ↪ registers, photocopy machines, and microfilm machines. Similar categories include 5021 (Office
    ↪ Furniture), 5045 (Computer Equipment), and 5943 (Stationery Stores).
Please provide the embedding of MCC 5044."
```

Listing 1: Enriched MCC embedding prompt with contextual information

LLM Embeddings. We evaluate several open-source LLMs and extract sentence embeddings as the final-layer representation at the last non-padding token. We store embeddings per field (location, MCC, merchant) for reuse and cold-start coverage.

Initialization and Training. We initialize categorical embedding tables with the LLM-derived vectors, providing (i) immediate semantic structure, (ii) faster convergence, and (iii) improved generalization to unseen but related categories. Fine-tuning uses a multi-task objective (e.g., transaction metrics and prediction), balancing semantic richness with task accuracy while remaining efficient production deployment.

4 Experiments

In this section, we present the results of comprehensive experiments to evaluate the effectiveness of our LLM-based semantic initialization framework across multiple transaction understanding tasks.

Table 2: Transaction Metrics Assessment: Relative Improvement (RI) across different embedding initialization strategies

Model	MCC + Merchant	Location	All Fields
Vanilla	1.00	1.00	1.00
Llama2-7b	-0.40%	+2.85%	+3.72%
Llama2-13b	+2.66%	+1.77%	+2.92%
Llama3-8b	+0.37%	+2.78%	+3.32%
Mistral-7b	+0.83%	+2.89%	+3.93%

We use four different LLMs for embedding initialization: LLama2-7b and LLama2-13b Touvron et al. [2023], Llama3-8b Grattafiori et al. [2024], and Mistral-7b Jiang et al. [2024a].

4.1 Experimental Setup

Dataset. We conducted experiments using one billion transaction records from January 2022 to December 2023. Transactions from the first 20 months are used for model training, the 21st month serves as the validation data, and transaction from the final 3 months is used as the test data.

Evaluation Tasks. We assess our approach through five tasks: (1) *Next Amount*: regression for transaction amount forecasting; (2) *Next MCC*: multi-class classification of MCC; (3) *Next Location*: city-level location prediction; (4) *Next Merchant*: fine-grained merchant classification; (5) *Transaction Metrics Assessment*: binary classification for anomaly detection using proprietary metrics.

Metrics. We report mean absolute error (MAE) and symmetric mean absolute percentage error (sMAPE) for regression, accuracy and F1-score for classification. For the proprietary task, we use Relative Improvement (RI) compared to the deployed baseline due to confidentiality constraints.

Baseline. We compare against a vanilla foundation model that uses randomly initialized, traditional ID-based categorical representations without semantic embeddings.

4.2 Results and Analysis

Table 1 shows comprehensive results across LLM architectures and embedding initialization strategies. Our key findings are: (1) *Initialization Strategy Impact*: Single-field strategies (MCC/Merchant-only) show specialized benefits, while combined approaches achieve broader improvements. (2) *Task-Specific Performance*: MCC/merchant prediction shows consistent gains (82%/100% of configurations), while amount prediction benefits from geographic and holistic semantic understanding. Location prediction shows modest improvements due to task simplicity. (3) *Architecture Analysis*: Llama3-8b demonstrates superior versatility across initialization strategies and tasks.

Table 2 presents RI results for transaction metrics assessment, where nearly all strategies outperform the baseline, demonstrating real-world effectiveness.

The results validate that LLM-generated semantic embeddings effectively capture merchant relationships and transaction patterns, with performance gains varying by task complexity and initialization strategy.

5 Conclusion

In this work, we addressed the limitations of traditional transaction analysis models that rely on index-based representations for categorical fields, leading to a loss of valuable semantic information. By proposing a hybrid framework that integrates LLM-generated embeddings as semantic initializations for lightweight transaction models, we achieve a balance between interpretability and operational efficiency, making it suitable for real-time financial applications. Our approach further leverages multi-source data fusion and a one-word constraint principle to ensure consistency and robustness in embedding generation, regardless of the underlying LLM architecture. Through systematic noise filtering and context-aware data enrichment, we enhance data quality and model reliability. Empirical results on large-scale transaction datasets validate the effectiveness of our framework, demonstrating significant improvements across multiple transaction understanding tasks.

References

Retail financial services — merchant category codes. International Standard ISO 18245:2023, International Organization for Standardization, 2023. URL `https://www.iso.org/standard/79450.html`.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sasaki, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yingzhi He, Xiaohao Liu, An Zhang, Yunshan Ma, and Tat-Seng Chua. Llm2rec: Large language models are powerful embedding models for sequential recommendation. *arXiv preprint arXiv:2506.21579*, 2025.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*, 2022.

Chengkai Huang, Tong Yu, Kaige Xie, Shuai Zhang, Lina Yao, and Julian McAuley. Foundation models for recommender systems: A survey and new perspectives. *arXiv preprint arXiv:2402.11143*, 2024.

AQ Jiang, A Sablayrolles, A Mensch, C Bamford, DS Chaplot, Ddl Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. Mistral 7b. arxiv 2023. *arXiv preprint arXiv:2310.06825*, 2024a.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. Scaling sentence embeddings with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3182–3196, 2024b.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*, 2024.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative filtering recommender systems. In *The adaptive web: methods and strategies of web personalization*, pages 291–324. Springer, 2007.

Piotr Skalski, David Sutton, Stuart Burrell, Iker Perez, and Jason Wong. Towards a foundation purchasing model: Pretrained generative autoregression on transaction sequences. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 141–149, 2023.

Aofeng Su, Aowen Wang, Chao Ye, Chen Zhou, Ga Zhang, Gang Chen, Guangcheng Zhu, Haobo Wang, Haokai Xu, Hao Chen, et al. Tablegpt2: A large multimodal model with tabular data integration. *arXiv preprint arXiv:2411.02059*, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Boris Van Breugel and Mihaela Van Der Schaar. Why tabular foundation models should be a research priority. *arXiv preprint arXiv:2405.01147*, 2024.

Yazheng Yang, Yuqi Wang, Guang Liu, Ledell Wu, and Qi Liu. Unitabe: A universal pretraining protocol for tabular foundation model in data science. *arXiv preprint arXiv:2307.09249*, 2023.

Chin-Chia Michael Yeh, Uday Singh Saini, Xin Dai, Xiran Fan, Shubham Jain, Yujie Fan, Jiarui Sun, Junpeng Wang, Menghai Pan, Yingtong Dou, Yuzhong Chen, Vineeth Rakesh, Liang Wang, Yan Zheng, and Mahashweta Das. Treasure: A transformer-based foundation model for high-volume transaction understanding. *in submission*, 2025.

Han Zhang, Xumeng Wen, Shun Zheng, Wei Xu, and Jiang Bian. Towards foundation models for learning on tabular data. 2023.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

A Preprocessing Details and Prompt Templates

This appendix details multi-source fusion, noise filtering, and prompt templates. We generate semantically rich embeddings for three primary categorical fields: location information, MCC, and merchant identifiers. Each field type requires specialized prompt engineering to capture domain-specific semantic relationships.

A.1 Location Embedding Generation

Location prompts incorporate geographical, economic, and regulatory context relevant to financial transactions. We design prompts that capture economic indicators, and regional financial attibutes.

```
Input: "UNITED STATES OF AMERICA, New York"
Prompt: "Represent the following location in the context of financial transactions, and economic
↪ indicators: New York, USA.
Consider state-specific economic trends, population demographics, major industries, and financial
↪ regulations."
```

Listing 2: Location embedding prompt example for state-level data

A.2 MCC Embedding Generation

MCC prompts leverage official category descriptions and business characteristics to generate semantically meaningful representations. We provide both basic and enriched versions depending on available contextual information. For a comprehensive MCC representations, we incorporate detailed business descriptions, included categories, and similar merchant types:

```
Input: MCC "5044"
Prompt: "The MCC 5044, titled 'Photographic, Photocopy, Microfilm Equipment and Supplies', serves
↪ business-to-business distributors of office and photographic equipment including film, cash
↪ registers, photocopy machines, and microfilm machines. Similar categories include 5021 (Office
↪ Furniture), 5045 (Computer Equipment), and 5943 (Stationery Stores).
Please provide the embedding of MCC 5044."
```

Listing 3: Enriched MCC embedding prompt with contextual information

A.3 Merchant Embedding Generation

Merchant prompts combine location, MCC category, and business name information to create comprehensive merchant representations:

```
Input: "365 MARKET 888 432-3299"
Prompt: "The merchant '365 MARKET 888 432-3299' is located in Troy, Michigan, USA. It belongs to MCC
↪ category 5814 'Fast Food Restaurants', which serves prepared food and beverages for on-premises
↪ or carry-out consumption.
Please provide the merchant embedding."
```

Listing 4: Merchant embedding prompt combining location and category context

Limitations

While our framework demonstrates significant improvements across multiple transaction understanding tasks, several limitations warrant acknowledgment and suggest directions for future research.

Prompt Engineering Sophistication: Our Explicit One-word Limitation Prompt Design, while effective for consistency, represents a relatively simple approach to prompt engineering. More sophisticated prompt optimization techniques and domain-specific fine-tuning of LLMs for financial contexts could potentially enhance semantic representation quality.

Model Coverage: Our evaluation focuses on established LLM architectures and does not include the most recent state-of-the-art models or specialized sentence embedding models such as NV-Embed Lee et al. [2024] and Qwen3-embedding Zhang et al. [2025]. Testing with these advanced models could

potentially yield further performance improvements and provide additional insights into architectural suitability for financial domain applications.

Categorical Field Scope: Our framework currently addresses MCC, merchant, and location embeddings, but does not extend to other potentially valuable categorical fields such as transaction channels, payment methods, or temporal patterns. The generalizability of our approach to these additional fields remains to be validated.

Static Embedding Approach: The current framework generates embeddings offline and does not account for evolving merchant characteristics, seasonal business patterns, or dynamic market conditions. This static approach may limit the framework's ability to capture temporal semantic changes in transaction contexts.

These limitations highlight opportunities for future research while acknowledging the scope and constraints of our current contribution to LLM-based semantic enhancement in financial transaction understanding.