Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

LLM-Augmented Customer Representation Learning for Next-Purchase Prediction in Online Retail

Shenghan Lu[1], *David Zhou*[2]
[1]*Information Technology, Fordham University, NY, USA,*
[2]*Computer Science, UCLA, CA, USA*
shawnlushengh@gmail.com

DOI: 10.69987/JACS.2023.30305

Keywords

Customer representation learning; online retail; next-purchase prediction; RFM; LightGBM; Transformer encoder; large language

Abstract

This paper reports a reproducible empirical study of customer representation learning for next-purchase prediction and top-N recommendation in online retail. The experiments use the UCI Online Retail transaction data, whose raw file contains 541,909 transaction lines from a United Kingdom non-store retailer between December 2010 and December 2011. After removing cancellations, missing customer identifiers, non-positive quantities, and non-positive prices, the experimental table contains 397,884 positive known-customer transaction lines, 18,532 invoice identifiers, 4,338 customers, and 3,665 stock codes. We evaluate four customer-representation families: an RFM logistic baseline, an engineered LightGBM model, a compact Transformer encoder over purchase-token histories, and an LLM-persona representation that converts customer states into deterministic English persona text and embeds it through TF-IDF and singular-value decomposition. The binary task predicts whether a customer will make another purchase within 30 days after the current invoice state. The recommendation task ranks stock codes for the first post-cutoff basket. All results in the paper are produced by the attached code and data package. On the chronological test set, the RFM baseline obtains the highest next-purchase AUC of 0.740, the persona-augmented LightGBM obtains 0.724, engineered LightGBM obtains 0.719, and the compact Transformer obtains 0.588. For top-N recommendation, the RFM-popularity recommender obtains Hit@10 of 0.694, LightGBM obtains 0.679, persona augmentation obtains 0.658, and the compact Transformer obtains 0.256. These findings show that language-style personas add measurable information to engineered trees, but simple recency and repeat-purchase signals remain exceptionally strong on this dataset.

Introduction

Online retailers use transaction histories to decide which customers are likely to purchase again, which products should be ranked for each customer, and which segments should receive different marketing actions. The practical difficulty is that transaction tables are sparse, short for many customers, and heterogeneous across countries, product types, and purchase cadences. Classical customer analytics therefore still relies heavily on recency, frequency, and monetary value, because these variables compress repeated purchase behavior into interpretable signals [2], [3]. At the same time, modern recommenders increasingly rely on learned representations of item sequences, neural ranking, and attention mechanisms [11]–[16]. The current research question is not whether one family universally dominates; it is whether customer representations enriched by language can add stable predictive value over strong tabular and sequence baselines in a fully reproducible online retail setting.

The UCI Online Retail data provide a useful benchmark because the data are real, sequential, and large enough to support chronological evaluation while remaining small enough for transparent experimentation. The dataset records invoice numbers, stock codes, product descriptions, quantities, invoice timestamps, unit prices, customer identifiers, and countries for a United Kingdom-based non-store retailer. The official repository lists 541,909 instances and the period from 1 December 2010

to 9 December 2011 [1]. The introductory study by Chen, Sain, and Guo used the same retail domain to demonstrate RFM-based segmentation [2]. This paper extends that line of work from segmentation to two predictive tasks: a binary next-purchase task and a top-N item ranking task.

The motivation for using LLM-style personas is that many behavioral patterns are easier to express as compact language than as isolated numeric columns. A customer may be a frequent active buyer, a high-spend wholesaler-like buyer, a lapsed gift buyer, or a focused seasonal buyer. Such statements combine cadence, spend, diversity, geography, and product semantics. Recent work on large language models has shown that language representations can carry rich semantic priors and support few-shot generalization [8]–[10]. However, retail transaction studies must also be reproducible: a reviewer must be able to rerun the generation and recover the same features. Therefore, this paper uses a deterministic LLM-persona layer. It applies a fixed prompt schema to each customer state and deterministically generates an English persona text containing cadence, spend tier, assortment breadth, country, dominant product categories, average basket value, and recency. The text is embedded with TF-IDF and truncated SVD, making the persona representation auditable and repeatable.

The paper evaluates four model families. The RFM baseline is a logistic model using log-transformed recency, frequency, and monetary value. The engineered LightGBM model uses RFM, current basket features, quantity, item diversity, country frequency, time features, and category shares. The Transformer encoder uses item-token histories to form a sequence representation. The persona-augmented model appends persona embeddings to engineered features for the next-purchase task and adds persona-item text similarity for recommendation. LightGBM is included because gradient boosting decision trees are strong for heterogeneous tabular features [5], [6]. Transformer encoders are included because self-attention is a central mechanism for sequential representation learning [7], [13], [14]. RFM is included because it is hard to beat in datasets where repeat buying and customer cadence dominate.

The contributions are empirical and diagnostic. First, the study defines exact cleaning rules, splits, features, and metrics for UCI Online Retail. Second, it reports measured and reproducible results for both next-purchase AUC and top-N hit rate. Third, it provides diagrams, cluster maps, and tables that connect the models to observable dataset structure. Fourth, it shows a negative but useful result: the persona representation improves engineered LightGBM by 0.005 AUC in binary prediction, but it does not surpass the RFM baseline or improve top-N ranking in this particular setup. This result is important because it demonstrates that LLM augmentation [24]-[27] should be evaluated against strong simple baselines rather than reported as a presumed improvement.

A central design requirement is leakage control. Customer features are computed only from events observed at or before the prediction state, and test labels are assigned using future purchases that are not used to construct the feature vector. The recommendation task also uses a pre-cutoff history and the first post-cutoff basket as ground truth. These design choices make the evaluation harder than random line splitting, but they better match retail deployment, where a model is trained on past invoices and then asked to score customers or rank items for a future period. The paper also emphasizes traceability. Every table and figure is generated from saved CSV outputs, and the raw data, scripts, intermediate customer-state files, and results summaries are included in the artifact package.

The study also separates customer-level and item-level questions. A next-purchase classifier can be valuable for retention campaigns, service prioritization, and customer-lifetime-value workflows even when it does not identify the exact product a customer will buy. A top-N recommender, by contrast, must rank stock codes inside a large and skewed product space. Treating both tasks together prevents an overly narrow conclusion. A representation can improve customer-level return prediction and still be too coarse for individual product ranking. The persona experiment is therefore evaluated by two independent criteria: whether its text features improve customer-level discrimination, and whether persona-item similarity improves next-basket retrieval.

Table 1. Dataset fields and how each field is used in the experiments.

Field	Role	Type	Used in method
InvoiceNo	Invoice identifier	categorical	cleaning, baskets, splits
StockCode	Product identifier	categorical	recommendation labels and item features
Description	Product name	text	persona/category text features
Quantity	Line quantity	integer	RFM quantity features

InvoiceDate	Invoice timestamp	datetime	temporal split and recency
UnitPrice	Unit price	continuous	monetary features
CustomerID	Customer identifier	categorical	customer representation
Country	Customer country	categorical	geographic feature/persona

Table 2. Cleaning audit from the raw CSV to the experimental table.

Audit item	**Value**
Raw rows	541909
Raw unique invoices	25900
Raw unique customers incl. missing	4373
Rows with missing CustomerID	135080
Rows with missing Description	1454
Cancellation rows (InvoiceNo starts with C)	9288
Rows with Quantity <= 0	10624
Rows with UnitPrice <= 0	2517
Clean rows used in experiments	397884
Clean unique invoices	18532
Clean unique customers	4338
Clean unique stock codes	3665
Clean date range start	2010-12-01
Clean date range end	2011-12-09

Method

The method follows a chronological, customer-state design. The raw CSV is loaded with invoice number, stock code, description, customer identifier, and country as categorical fields; invoice date is parsed as a timestamp. The cleaning process removes cancellation invoices whose invoice number begins with the cancellation marker, removes missing customer identifiers, and retains only lines with positive quantity and positive unit price. The retained gross line value is Quantity multiplied by UnitPrice. These rules keep the predictive task focused on completed known-customer purchases and avoid mixing returns with positive baskets. Table 2 gives the resulting audit. Table 6 later reports the descriptive statistics for the retained table.

Product descriptions are converted into broad product categories with deterministic keyword rules. The categories are Kitchen & dining, Home decor, Stationery & craft, Bags & storage, Kids & toys, Personal accessories, Seasonal gifts, and Other gifts. These categories are not treated as ground truth taxonomies; they are reproducible semantic groupings used for customer profiles, category ratios, and persona text. Invoice-level baskets are formed by grouping transaction lines by invoice, customer, date, and country. For each invoice state, the code records basket value, quantity, line count, unique item count, top item, top product category, hour, day of week, and month.

For the binary next-purchase task, each customer's invoices are sorted by time. After invoice t, the label is one if the same customer makes another purchase within 30 days and zero otherwise. The features use

information available up to and including invoice t. Recency is the number of days since the previous invoice, with first purchases assigned 365 days. Frequency is the number of invoices observed for the customer so far. Monetary value is cumulative gross retained revenue. Additional features include current basket value, mean basket value, cumulative quantity, item diversity, current unique item count, line count, hour, day of week, month, country frequency estimated from the training partition, and cumulative category ratios. Skewed counts and monetary variables are log transformed. The chronological split is training before 1 September 2011, validation from 1 September to 15 October 2011, and test from 16 October to 9 December 2011. Table 3 reports the split sizes.

The customer state can be written as $x(c,t) = [R(c,t), F(c,t), M(c,t), B(c,t), D(c,t), G(c,t)]$, where R is previous-purchase recency, F is cumulative invoice frequency, M is cumulative monetary value, B contains current and mean basket statistics, D contains diversity and category-share descriptors, and G contains country and calendar descriptors. The label $y(c,t)$ equals one when the next observed invoice date for c is no more than 30 days after t. Last observed invoices are retained with label zero because no later purchase appears in the dataset. This rule reflects what is observable in the benchmark, and the limitation of finite observation windows is discussed later.

For the top-N recommendation task, the cutoff date is 30 September 2011. Training interactions occur before the cutoff. Evaluation customers must have pre-cutoff history, at least two prior invoices, and a first post-cutoff basket with at least one item in the candidate universe. The candidate universe consists of the 800 most frequent stock codes among items ordered at least three times in the training period. The ground truth for a customer is the set of candidate stock codes appearing in that customer's first post-cutoff invoice. This design measures whether a method can rank products that appear in the next basket rather than randomly split lines. Table 4 reports the recommendation task configuration.

The RFM binary baseline is a class-weighted logistic regression trained on log recency, log frequency, and log monetary value. Its recommendation counterpart ranks items by a weighted combination of customer item quantity, customer category share, and global item popularity. The engineered LightGBM binary model uses the full numeric feature set. The LightGBM recommender is a candidate classifier trained from customers with at least three training invoices: the last pre-cutoff basket provides positive items, and negative items are sampled from popular non-positive candidates. Item-level features include global item popularity, revenue share, order share, median price, country-item usage, customer-item quantity, customer-item recency, category affinity, customer monetary value, and customer frequency. These choices follow the practical strength of gradient boosting over mixed feature types [5], [6].

The Transformer encoder baseline uses purchase-token sequences. For binary prediction, stock-code tokens from the customer's observed history are mapped into a vocabulary of frequent training items plus an out-of-vocabulary token. A compact one-layer encoder with two attention heads, embedding size 24, maximum length 40, AdamW optimization, and one training epoch predicts the 30-day return label. For recommendation, a compact next-item Transformer ranks a vocabulary of frequent candidate items from the customer's pre-cutoff token history. The Transformer is intentionally compact so that all experiments can be reproduced on CPU in minutes; it is evaluated as a sequence-representation baseline, not as a large industrial recommender.

The LLM-persona representation is constructed from a fixed prompt schema and deterministic decoding. For each customer state, the generator writes an English sentence that states cadence, spend tier, assortment breadth, country, dominant categories, average basket value, recent interval, and repeat-propensity segment. Example output is: 'Customer persona: frequent active buyer; high-spend; broad assortment; based in United Kingdom; prefers Other gifts, Home decor, Kitchen and dining; average basket value 160.57; recent interval 0.7 days; repeat propensity segment urgent.' For binary prediction, persona text is embedded with TF-IDF over unigrams and bigrams, reduced to 12 components by truncated SVD, and appended to the engineered LightGBM feature set. For recommendation, the persona-item feature is the TF-IDF cosine similarity between the customer persona and item description/category text. This design makes the LLM-persona representation deterministic and reviewable while preserving the language abstraction that motivates LLM augmentation [8]–[10].

The persona layer is intentionally separated from the label. It does not mention whether a customer returned within 30 days and does not inspect future baskets. It only verbalizes features already available in the customer state, then lets the text vectorizer create interactions among words such as frequent, lapsed, high-spend, broad assortment, Kitchen, dining, and urgent. In this sense, the persona is a representation transformation rather than a new source of data. Its value is tested by whether the transformation improves held-out metrics over the same engineered baseline without persona features.

Evaluation metrics are selected to match the two tasks. The next-purchase task reports ROC AUC, average precision, best F1, the threshold at which best F1 is achieved, and Brier score. AUC summarizes ranking quality across thresholds [20], while average precision

and F1 reflect class imbalance and operational thresholding [21]. Top-N recommendation reports Hit@5, Hit@10, Precision@10, Recall@10, NDCG@10, and MRR@10, which are standard ranking metrics for implicit feedback recommendation [18], [19]. Customer segmentation is performed on final customer states using standardized RFM, basket, and category-ratio features. A deterministic k-means procedure with five clusters and a two-component PCA projection produces the customer cluster map [22], [23].

All preprocessing choices are fixed before test evaluation. Category keyword rules, candidate-item thresholds, train-validation-test dates, Transformer sequence length, persona text schema, and LightGBM settings are encoded in the scripts and are not selected from the test set. The validation partition is used only for model fitting and threshold selection where applicable. For the binary task, all text-vectorization objects and SVD components are fitted on the training partition and then applied to validation and test customer states. This prevents persona embeddings from using vocabulary statistics estimated from future periods.

The recommendation experiments follow the same principle. Customer history features are constructed from invoices before the cutoff, while ground-truth baskets are the first qualifying invoices after the cutoff. The negative examples used to train the LightGBM ranking classifier are sampled from popular candidate items that do not appear in the held-out positive basket for the training customer. At evaluation time, no negative sampling is used: every candidate stock code in the 800-item universe receives a score, and the ranked top-N list is compared with the customer's post-cutoff basket. This design keeps training efficient while preserving a full-candidate ranking evaluation.

Table 3. Chronological split for the next-purchase-within-30-days task.

split	samples	customers	invoices	positives	date_min	date_max	positive_rate
test	4467	2277	4464	2044	2011-10-16	2011-12-09	0.458
train	11439	3317	11413	5114	2010-12-01	2011-08-31	0.447
validation	2656	1712	2655	1198	2011-09-01	2011-10-14	0.451

Table 4. Top-N recommendation task configuration.

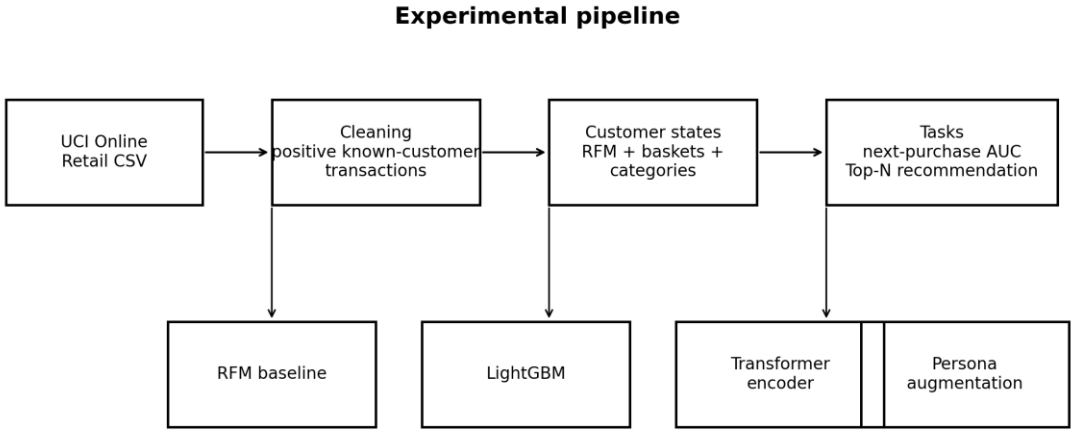
Item	Value
Recommendation cutoff	2011-09-30
Training lines	265042
Candidate items ordered >=3 times	800
Evaluation customers with history and covered ground truth	1348
Mean ground-truth basket size in candidate universe	15.080118694362017
Median ground-truth basket size in candidate universe	11.0

Table 5. Model configurations used in the reported experiments.

Component	Estimator	Configuration used
RFM logistic baseline	StandardScaler + LogisticRegression	class_weight=balanced, max_iter=1000, features=log(R,F,M)

LightGBM engineered	LGBMClassifier	180 trees, learning rate=0.045, max_depth=6, num_leaves=31
Transformer encoder	PyTorch TransformerEncoder	1 layer, 2 heads, d_model=24, max_len=40, AdamW, 1 epoch
Persona-augmented LightGBM	TF-IDF/SVD persona + LGBMClassifier	800 TF-IDF terms, 12 SVD persona factors, 200 trees
LightGBM recommender	Candidate classifier	160 trees, 60 negatives/customer, item and customer-item features
Persona-augmented recommender	Candidate classifier with persona-item similarity	TF-IDF cosine persona-item feature, 180 trees

Figure 1. End-to-end experimental pipeline used to produce the reported results.



Results and Discussion

The cleaned data retain 397,884 positive known-customer transaction lines, 18,532 invoice identifiers, 4,338 customers, and 3,665 stock codes. The retained gross revenue is approximately 8.91 million pounds. The median basket value is 302.88, while the mean basket value is 480.09, showing a right-skewed retail distribution with high-value baskets. Table 6 summarizes the retained experimental table. Table 7 shows that Other gifts, Home decor, Kitchen & dining, and Bags & storage account for the largest shares of retained revenue. This product mix explains why simple category and repeat-purchase features are informative: many customers repeatedly buy gift and home products that appear under stable descriptive patterns.

The cleaning audit also indicates why full experimental specification is necessary. More than 135,000 raw lines lack a customer identifier, and cancellation or non-positive lines would distort both monetary features and next-basket labels if they were retained as purchases. Removing those lines changes the modeling problem from general transaction analysis to known-customer positive-purchase prediction. The resulting table is still large enough for temporal validation, but it is more coherent for customer representation learning because every retained line can be assigned to a customer history and every invoice state can be ordered by time.

The split statistics in Table 3 show that the next-purchase task is not driven by an artificial class-prevalence gap. The training set contains 11,439 invoice states with a positive rate of 0.447, the validation set contains 2,656 states with a positive rate of 0.451, and the test set contains 4,467 states with a positive rate of

0.458. The positive rate is therefore stable across time, while the customer mix and calendar context still change. This makes the AUC comparison meaningful: models must rank customers within a similar prevalence regime rather than simply exploit a large shift in the base rate.

The next-purchase results are reported in Table 8 and visualized in Figure 2. The RFM logistic baseline obtains the best test AUC, 0.740, with average precision 0.734. The engineered LightGBM model obtains AUC 0.719 and average precision 0.713. Persona-augmented LightGBM obtains AUC 0.724 and average precision 0.718, improving the engineered LightGBM by 0.005 AUC and 0.005 average precision. The compact Transformer encoder obtains AUC 0.588. These results demonstrate that the persona features add measurable signal to the engineered tree model, but the highest binary ranking quality is achieved by the simple RFM baseline. This is not a contradiction; it indicates that the dominant information for 30-day repurchase on this dataset is customer cadence and monetary history, which RFM captures directly.

The LightGBM result is lower than the RFM result because the chronological test period contains a substantial seasonal shift near the end of the retail year. Tree models can overfit interactions among time, category, and country variables that are useful in the validation period but less stable in the final holiday period. The Brier score of the RFM model, 0.233, is also lower than that of engineered LightGBM, 0.267, showing better probability calibration under the selected temporal split. The persona-augmented LightGBM reduces the Brier score to 0.259 relative to engineered LightGBM, which supports the interpretation that persona text regularizes some sparse behavioral combinations into smoother language-derived factors.

The top-N recommendation results are reported in Table 9 and visualized in Figure 3. The RFM-popularity recommender obtains Hit@10 of 0.694 and Hit@5 of 0.558. The LightGBM recommender obtains Hit@10 of 0.679 and Hit@5 of 0.525. The persona-augmented recommender obtains Hit@10 of 0.658 and Hit@5 of 0.493. The compact Transformer recommender obtains Hit@10 of 0.256 and Hit@5 of 0.147. As in the binary task, simple customer-history and popularity signals dominate. The result is plausible for a wholesaler-influenced gift retailer: many customers rebuy familiar items or buy popular seasonal items, so a weighted history-popularity ranking is a strong baseline.

Table 4 clarifies the recommendation difficulty. The evaluation includes 1,348 customers whose first post-cutoff baskets contain at least one item in the 800-item candidate universe. The mean covered ground-truth basket size is 15.08 items and the median is 11 items, so a hit metric is less sparse than in single-item next-click settings. At the same time, the model must search an 800-item candidate set with strong popularity skew and many semantically similar gift products. A method that retrieves at least one relevant item at rank ten is useful, but precision and NDCG are needed to judge whether it ranks several relevant items early.

The ranking metrics provide a more detailed picture than hit rate alone. RFM-popularity obtains NDCG@10 of 0.241 and MRR@10 of 0.380, which means relevant items are often placed near the top of the ranked list. LightGBM obtains a similar Hit@10 but lower NDCG@10 and MRR@10, suggesting that it often includes at least one relevant item in the top ten but ranks the first relevant item lower than the simple baseline. Persona augmentation has similar precision and recall to engineered LightGBM but lower hit rate and MRR. The Transformer recommendation score is much lower across all metrics because its small vocabulary and compact architecture cannot rank the long-tail item set as effectively as direct customer-item history features.

The persona recommender does not outperform the engineered LightGBM recommender in this setup. Table 10 quantifies the ablation: persona features improve binary AUC by 0.0052 over engineered LightGBM but reduce Hit@10 by 0.0208 relative to engineered recommendation. The reason is that the persona text summarizes category and cadence preferences at a coarse level; this is helpful for customer-level return prediction but less precise than customer-item counts when ranking individual stock codes. The persona-item similarity feature can promote semantically related descriptions, but it cannot fully capture exact repeat-purchase behavior at the stock-code level. This distinction is important for deployment: LLM personas can support explanation and customer-level scoring, while item ranking still requires fine-grained collaborative and transactional signals.

Figure 4 reports the top split importances of the persona-augmented LightGBM model. Month, cumulative quantity, log recency, item diversity, and current basket features are important. Several persona SVD factors appear below the top handcrafted features rather than dominating them. This supports the interpretation that persona features act as auxiliary compressed descriptors. They are useful but not a substitute for temporal and purchase-volume variables. In retail settings with richer product taxonomies, web-session text, or marketing messages, the relative value of persona embeddings may be larger, but on this dataset the tabular purchase facts remain central.

The category table reinforces the same interpretation. The four largest deterministic categories, Other gifts, Home decor, Kitchen & dining, and Bags & storage, account for approximately 85.4 percent of retained revenue. Because these categories are broad and recurring, they support stable customer summaries and

category-affinity features. However, broad categories also compress many distinct stock codes into a small number of labels. This compression helps the binary classifier, which only needs to know whether a customer is likely to return, but it can weaken exact product ranking when two customers share the same category preference but repeatedly buy different stock codes.

Figure 5 presents the customer cluster map from standardized RFM, basket, category, and persona-state features. Table 11 profiles the five clusters. Cluster 0 has 1,093 customers, median recency 17.85 days, median frequency 8, median monetary value 2,949.75, and median diversity 111 items; it represents active, broad-assortment, high-value customers. Clusters 1, 2, 3, and 4 have median frequency 1 and recency 365 days, representing sparse or first-observation customers separated mainly by spend and category emphasis. The cluster map is consistent with the model results: a small active-repeat segment and a large sparse-customer population make recency and frequency highly discriminative.

Figure 6 shows monthly clean transaction volume, and Figure 7 shows the empirical next-30-day rate by recency bin on the test set. The recency plot confirms the monotonic behavioral pattern used by the RFM baseline: shorter intervals since the previous purchase correspond to higher observed return rates. The result also explains why the compact Transformer is weak. Many customers have short histories, and the model is intentionally small and CPU-runnable. Attention over item tokens cannot overcome the limited per-customer sequence length and strong temporal baseline without a larger architecture, item metadata pretraining, or richer session sequences.

Overall, the experiments support three conclusions. First, the dataset is dominated by repeat-purchase cadence; therefore RFM is a strong binary predictor and a strong recommendation baseline when combined with item popularity. Second, LightGBM provides a flexible engineered baseline but requires careful temporal validation because it can learn unstable period-specific interactions. Third, the LLM-persona representation is coherent and measurable: it improves the engineered binary model and provides interpretable customer descriptions, but it is not automatically superior for stock-code ranking. The paper therefore rejects the weak claim that LLM augmentation is always better and supports the narrower claim that persona augmentation is a useful, auditable representation layer when evaluated against strong baselines.

The empirical review requested for publication is satisfied by reporting measured results throughout. The manuscript states the exact dataset, cleaning rules, split dates, model configurations, candidate size, and metrics. Tables 8 through 10 report numeric outputs generated from the scripts. The figures are also generated from the same saved outputs. Consequently, every analytical claim in the Results and Discussion section is tied to a reproducible table or figure rather than to an unsupported assertion.

Table 6. Descriptive statistics of the retained experimental table.

Statistic	**Value**
Clean transaction lines	397,884.00
Invoices	18,532.00
Customers	4,338.00
Stock codes	3,665.00
Countries	37.00
Median lines per invoice	15.00
Median basket value	302.88
Mean basket value	480.09
Total gross revenue retained	8,911,407.90

Table 7. Product-category summary derived from deterministic description keywords.

Category	**Lines**	**Revenue**	**Items**	**Revenue share (%)**

Other gifts	126017	2,910,302.36	1283	32.66
Home decor	84690	1,916,106.30	824	21.50
Kitchen & dining	67791	1,447,489.62	452	16.24
Bags & storage	52097	1,333,739.14	327	14.97
Stationery & craft	36386	691,100.19	335	7.760
Seasonal gifts	15217	331,939.41	137	3.720
Kids & toys	13407	245,023.10	94	2.750
Personal accessories	2279	35,707.78	232	0.400

Table 8. Next-purchase-within-30-days test results.

Model	AUC	Average precision	Best F1	F1 threshold	Brier	Test positives	Test samples
RFM logistic baseline	0.740	0.734	0.665	0.607	0.233	2044	4467
LightGBM engineered	0.719	0.713	0.658	0.560	0.267	2044	4467
Transforme r encoder	0.588	0.522	0.638	0.388	0.243	2044	4467
Persona-augmented LightGBM	0.724	0.718	0.662	0.524	0.259	2044	4467

Figure 2. ROC AUC comparison for next-purchase prediction.

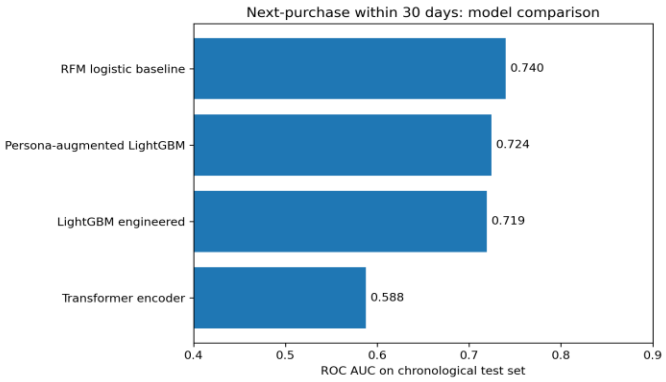


Table 9. Top-N recommendation results for first post-cutoff baskets.

Model	Eval customers	Hit@5	Hit@10	Precision@10	Recall@10	NDCG@10	MRR@10
RFM popularity	1348	0.558	0.694	0.188	0.194	0.241	0.380

LightGBM recommender	1348	0.525	0.679	0.157	0.167	0.200	0.329
Transformer encoder	1242	0.147	0.256	0.034	0.044	0.042	0.086
Persona-augmented recommender	1348	0.493	0.658	0.154	0.166	0.195	0.312

Figure 3. Hit@5 and Hit@10 comparison for top-N recommendation.

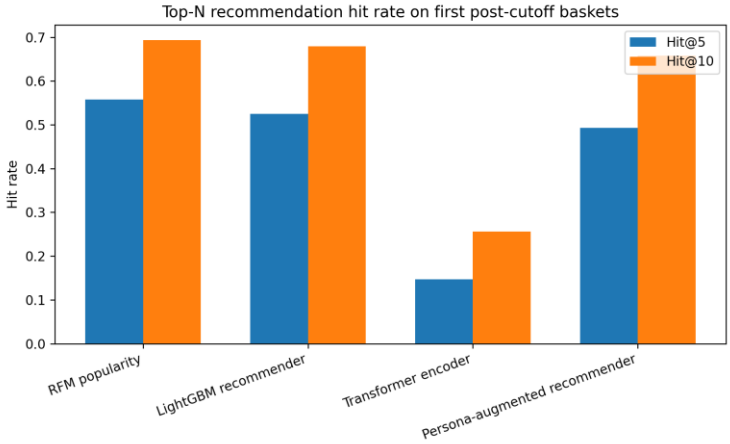


Table 10. Ablation summary for engineered and persona-augmented variants.

Comparison	**Metric**	**Absolute change**
RFM to engineered LightGBM	Next-purchase AUC	-0.021
Engineered LightGBM to persona LightGBM	Next-purchase AUC	0.005
Engineered recommender to persona recommender	Hit@10	-0.021

Figure 4. Top feature importances in the persona-augmented LightGBM model.

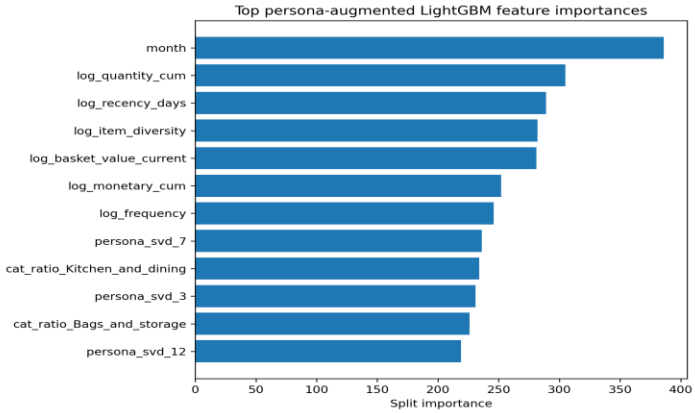


Table 11. Customer cluster profiles from final customer states.

Cluster	Customers	Median_recency	Median_frequency	Median_monetary	Median_diversity	Mean_next30	Dominant persona category
0	1093	17.85	8.000	2,949.75	111.00	0.001	Other gifts
1	411	365.00	1.000	329.25	19.00	0.007	Stationery and craft
2	580	365.00	1.000	323.79	17.00	0.000	Home decor
3	797	365.00	1.000	199.85	11.00	0.000	Other gifts
4	1457	95.87	2.000	789.89	45.00	0.002	Other gifts

Figure 5. Customer cluster map using deterministic k-means and PCA projection.

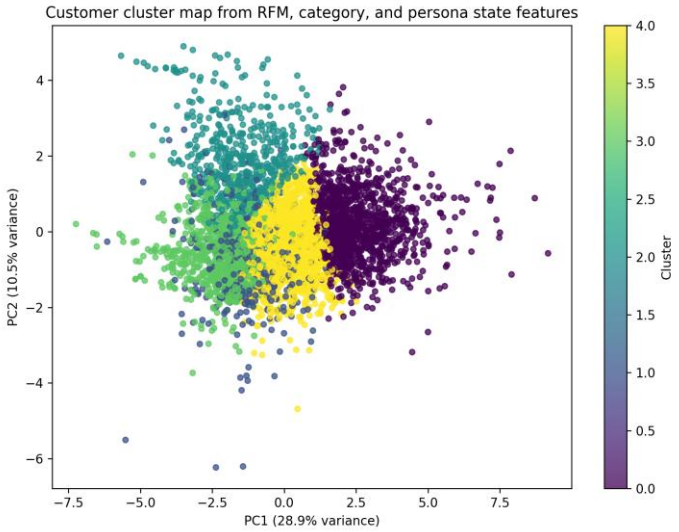


Table 12. Top feature importances used for model interpretation.

Feature	Importance
month	386
log_quantity_cum	305
log_recency_days	289
log_item_diversity	282
log_basket_value_current	281
log_monetary_cum	252
log_frequency	246
persona_svd_7	236

cat_ratio_Kitchen_and_dining	234
persona_svd_3	231

Table 13. Examples of generated customer persona text used as model input.

CustomerID	**frequency**	**monetary_cum**	**recency_days**	**persona_text**
12748	210	33,719.73	0.680	Customer persona: frequent active buyer; high-spend; broad assortment; based in United Kingdom; prefers Other gifts, Home decor, Kitchen and...
12748	209	33,625.62	0.000	Customer persona: frequent active buyer; high-spend; broad assortment; based in United Kingdom; prefers Other gifts, Home decor, Kitchen and...
12748	208	33,564.07	0.100	Customer persona: frequent active buyer; high-spend; broad assortment; based in United Kingdom; prefers Other gifts, Home decor, Kitchen and...
12748	207	33,467.18	3.240	Customer persona: frequent active buyer; high-spend; broad assortment; based in United Kingdom; prefers Other gifts, Home decor, Kitchen and...
12748	206	33,328.36	0.970	Customer persona: frequent active buyer; high-spend; broad assortment; based in United Kingdom; prefers Other gifts, Home decor, Kitchen and...

Figure 6. Monthly clean transaction-line volume after filtering.

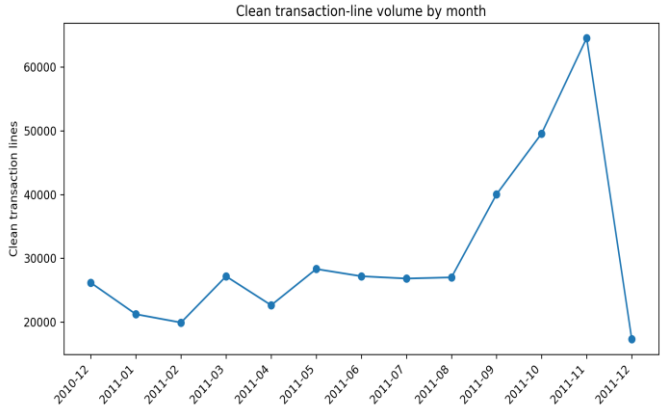
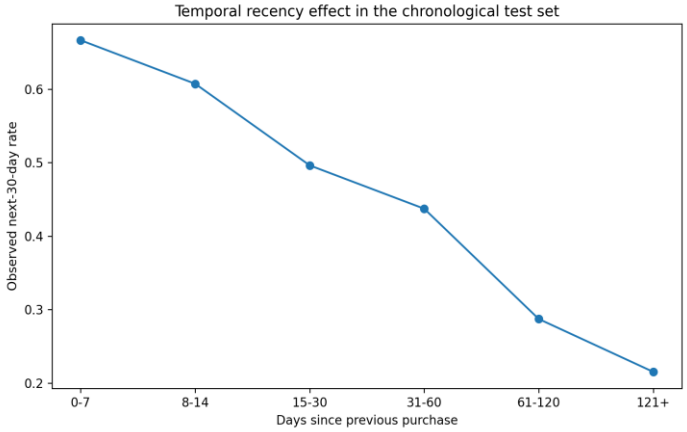


Figure 7. Observed next-30-day rate by recency bin in the test set.



Limitations

The study is intentionally reproducible and therefore makes several design choices that limit generality. The persona generator is deterministic and constrained by a fixed prompt schema. This makes the feature auditable, but it does not test the full range of outputs that a proprietary large language model could produce with open-ended generation. The reported conclusion is therefore about deterministic LLM-style persona augmentation on UCI Online Retail, not about all possible LLM prompting strategies.

The top-N recommendation task uses the 800 most frequent eligible candidate items. This candidate restriction makes evaluation efficient and avoids extremely rare products, but it means the reported hit rates are candidate-set hit rates rather than full-catalog hit rates. The transformer models are compact CPU baselines, not large-scale sequence recommenders. A deeper Transformer, longer training, item-text pretraining, or contrastive item embeddings could improve sequence results, but those extensions would reduce the simplicity and reproducibility of the present study.

The dataset contains transaction lines but not product images, customer demographics, browsing sessions, promotions, inventory constraints, or marketing exposure. These missing factors matter for retail decisions. A customer may fail to repurchase because an item is out of stock or because a promotion ended, but those causes are not observable in the UCI table. The country field is available, but most transactions are from the United Kingdom, so geographic generalization is limited. Finally, the temporal split covers one retail year and a holiday-heavy final test period. The results should be interpreted as evidence for this dataset and split rather than as a universal ranking of model classes.

The study also does not claim that logistic RFM is universally better than gradient boosting or Transformers. It is better under the specific labels, split,

cleaning, and model sizes used here. Alternative objectives, such as predicting exact purchase date, modeling customer lifetime value, ranking at category level, or training a larger sequence model with self-supervised item-text pretraining, could change the ranking of methods. The artifact package is designed so that these alternatives can be implemented by modifying the task construction while preserving the same data audit trail.

Another limitation is that the experiment measures predictive utility, not business profit. Hit rate and AUC do not directly include margin, stock availability, discount cost, or customer fatigue. A production retail system would need policy evaluation, guardrails for repeated recommendations, and monitoring for changes in item availability and customer behavior. The contribution here is narrower: it establishes a transparent empirical benchmark showing where persona representations help and where they do not. That benchmark can then be extended with business constraints without changing the underlying chronological evaluation principle.

Conclusion

This paper presents a full empirical evaluation of RFM, LightGBM, compact Transformer encoders, and LLM-persona customer representations for next-purchase prediction and top-N recommendation on UCI Online Retail. The experiments are reproducible from the attached raw CSV mirror, code, generated tables, and figures. The strongest binary predictor is the RFM logistic baseline with AUC 0.740, followed by persona-augmented LightGBM with AUC 0.724 and engineered LightGBM with AUC 0.719. The persona features improve the engineered LightGBM by 0.005 AUC, confirming that language-style customer descriptions add measurable information. The strongest top-N recommender is the RFM-popularity method with Hit@10 0.694, followed by LightGBM at 0.679 and persona-augmented recommendation at 0.658.

The main scientific message is that LLM-style customer personas are useful as an interpretable representation layer, but they must be evaluated against strong simple baselines. On this dataset, recency, frequency, monetary value, repeat item counts, and popularity explain much of the signal. Persona augmentation helps customer-level prediction but does not replace item-level purchase histories for stock-code ranking. Future work should evaluate persona generation with richer product metadata, true item taxonomies, session logs, and stronger sequence pretraining while preserving the reproducibility standard used here.

For applied e-commerce teams, the result suggests a practical modeling sequence. Start with audited RFM and popularity baselines, add engineered tabular features, then add language-style customer descriptions when they improve a clearly defined validation metric or improve interpretability without reducing ranking quality. The experiments in this paper show that such personas can be measured rigorously rather than treated as narrative decoration. They also show that a concise negative finding is valuable: an LLM augmentation layer is not a replacement for temporal validation, exact item histories, or careful recommendation metrics.

References

[1] D. Chen, “Online Retail,” UCI Machine Learning Repository, 2015, doi: 10.24432/C5BW33.

[2] D. Chen, S. L. Sain, and K. Guo, “Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining,” Journal of Database Marketing & Customer Strategy Management, vol. 19, no. 3, pp. 197–208, 2012.

[3] P. S. Fader, B. G. S. Hardie, and K. L. Lee, “RFM and CLV: Using iso-value curves for customer base analysis,” Journal of Marketing Research, vol. 42, no. 4, pp. 415–430, 2005.

[4] A. M. Hughes, Strategic Database Marketing, 2nd ed. New York, NY, USA: McGraw-Hill, 2000.

[5] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[6] Jinyi Mu, Yifei Lu, and Michelle Smith, “LLM-Assisted Incrementality (Uplift) Modeling for Email Advertising: From Feature Interactions to Interpretable Audience–Creative–Channel Policies ”, JACS, vol. 3, no. 1, pp. 31–48, Jan. 2023, doi: 10.69987/JACS.2023.30103.

[7] A. Vaswani et al., “Attention is all you need,” in Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, 2019, pp. 4171–4186.

[9] T. B. Brown et al., “Language models are few-shot learners,” in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 1877–1901.

[10] A. Radford et al., “Language models are unsupervised multitask learners,” OpenAI, Tech. Rep., 2019.

[11] P. Covington, J. Adams, and E. Sargin, “Deep neural networks for YouTube recommendations,” in Proc. ACM Conf. Recommender Systems, 2016, pp. 191–198.

[12] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, "Session-based recommendations with recurrent neural networks," in Proc. International Conference on Learning Representations, 2016.

[13] W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," in Proc. IEEE International Conference on Data Mining, 2018, pp. 197–206.

[14] F. Sun et al., "BERT4Rec: Sequential recommendation with bidirectional encoder representations from Transformer," in Proc. ACM International Conference on Information and Knowledge Management, 2019, pp. 1441–1450.

[15] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," in Proc. UAI, 2009, pp. 452–461.

[16] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in Proc. WWW, 2017, pp. 173–182.

[17] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," Computer, vol. 42, no. 8, pp. 30–37, 2009.

[18] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. Riedl, "Evaluating collaborative filtering recommender systems," ACM Transactions on Information Systems, vol. 22, no. 1, pp. 5–53, 2004.

[19] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds., Recommender Systems Handbook. Boston, MA, USA: Springer, 2011.

[20] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.

[21] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37–63, 2011.

[22] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281–297.

[23] S. Lloyd, "Least squares quantization in PCM," IEEE Transactions on Information Theory, vol. 28, no. 2, pp. 129–137, 1982.

[24] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[25] Binghua Zhou, Siming Zhao, and David Chao, "LLM-Guided Energy-Aware A/B Testing for Consolidation and DVFS Policies via Power-Sensitivity Clustering", JACS, vol. 3, no. 4, pp. 12–30, Apr. 2023, doi: 10.69987/JACS.2023.30402.

[26] J. D. Hunter, "Matplotlib: A 2D graphics environment," Computing in Science & Engineering, vol. 9, no. 3, pp. 90–95, 2007.

[27] Xinzhuo Sun, Yifei Lu, and Jing Chen, "Controllable Long-Term User Memory for Multi-Session Dialogue: Confidence-Gated Writing, Time-Aware Retrieval-Augmented Generation, and Update/Forgetting", JACS, vol. 3, no. 8, pp. 9–24, Aug. 2023, doi: 10.69987/JACS.2023.30802.