A Multi-Layer Hyper-Personalization Pipeline for Predicting Next User Actions Across Web, App, Email, and Social Channels

A Simulation-Based Feasibility Study

Kohei Arai
Saga University, 1 Honjo, Saga 840-8502 Japan

***Abstract*—Hyper-personalization systems that unify customer behavior across Web, App, Email, and social channels are increasingly central to digital business strategy, yet published evaluations rarely isolate which modeling component—collaborative filtering, sequential prediction, or behavioral clustering—actually drives predictive value, and rarely validate a full pipeline end-to-end before committing to production data collection. This study presents a complete hyper-personalization pipeline spanning unified behavior aggregation, recency/frequency/content feature engineering, item-based collaborative filtering, sequential next-action modeling (first-order Markov chains and LSTM networks), and K-means behavioral clustering, validated on a simulated multi-channel dataset (80,997 events, 600 users, 90 days) constructed with five known latent personas so that recovered structure could be checked against ground truth. The LSTM next-action model reached 34.5% validation accuracy against a 10.0% uniform-chance baseline and a 16.0% majority-class baseline, and outperformed the order-1 Markov chain by approximately 6 percentage points once evaluated on a matched held-out split. An ablation study further shows that 1) LSTM accuracy is insensitive to capacity across a 50-fold parameter range, suggesting that the bottleneck is behavioral signal rather than model size; 2) collaborative-filtering hit-rate@2 is strong despite weak absolute cosine similarities, indicating that relative ranking survives even when category-level content granularity limits similarity magnitude; and 3) clustering quality is substantially higher on a reduced recency/frequency/channel-mix feature set than on the full feature set including content preferences, implying that content preferences are better used downstream in recommendation than as clustering inputs. I report these findings, together with the pipeline's limitations on simulated versus real data, as a feasibility baseline for practitioners designing multi-channel personalization systems.**

Keywords*—Hyper-personalization; collaborative filtering; sequential recommendation; LSTM; Markov chain; customer segmentation; multi-channel analytics; ablation study

I. Introduction

Businesses increasingly engage customers across four or more digital channels simultaneously—a corporate website, a mobile application, email marketing, and social media—each generating its own behavioral signal in its own format and cadence. A customer who browses a product on the web, opens a follow-up email two days later, and comments on a related social post is, from the business's perspective, a single evolving intent; from the data's perspective, these are three disconnected event streams unless deliberately unified. Hyper-personalization—the practice of predicting what an individual user will want next from their full cross-channel behavioral history—depends on solving this unification problem before any predictive model can be trained at all.

A second, less discussed problem is architectural: personalization systems are typically built from one dominant technique—collaborative filtering for product recommendation, sequential models for session-based prediction, or clustering for segmentation—rather than as a combined pipeline whose components are individually validated and compared. Practitioners choosing between these approaches rarely have a controlled feasibility study to draw on, particularly one that isolates whether a more complex model such as an LSTM earns its complexity over simpler baselines such as a Markov chain or a majority-class predictor on their specific data shape.

This study addresses both problems directly. I construct a unified multi-channel behavioral pipeline and validate it on a simulated dataset engineered with known latent structure—five distinct user personas with different channel-activity rates and content preferences—so that every stage of the pipeline can be checked against a ground truth that real-world data never provides. My contributions are threefold.

A complete, reproducible hyper-personalization pipeline covering behavior aggregation, feature engineering, collaborative filtering, sequential modeling, and clustering, implemented and validated end-to-end.

A controlled comparison of majority-class, Markov-chain, and LSTM baselines for next-action prediction on matched held-out data, resolving an apparent parity that informal per-user inspection suggested but rigorous evaluation did not support.

An ablation study isolating the effect of the collaborative-filtering weighting scheme, LSTM capacity, and clustering feature composition, surfacing a counter-intuitive finding that a reduced feature set produces substantially better-separated clusters than the full engineered feature set.

The remainder of this study is organized as follows. Section II reviews related work in collaborative filtering, sequential recommendation, and customer segmentation. Section III describes the simulated dataset and the full pipeline architecture.

Section IV details the experimental setup and evaluation metrics. Section V reports the main results. Section VI presents the ablation study. Section VII discusses limitations, Section VIII discusses implications, and Section IX concludes.

II. Related Work

A. Collaborative Filtering

Collaborative filtering (CF) recommends items to a user based on patterns of co-engagement across a population, rather than content attributes of the items themselves. Sarwar et al. [1] introduced item-based CF, computing similarity between items from the co-rating patterns of users, which scales more favorably than user-based CF for large catalogs. Koren et al. [2] later formalized matrix-factorization approaches to CF, learning latent factors for users and items jointly, which became the dominant paradigm following the Netflix Prize. My collaborative-filtering component follows the item-based tradition of [1], computing cosine similarity between content categories from a user x content-type engagement matrix, chosen for interpretability and its established track record for the category-level granularity used here.

B. Sequential and Session-Based Recommendation

Markov-chain models capture the probability of a user's next action conditional on their most recent action(s), and remain a strong, fully interpretable baseline for next-item prediction; Rendle et al. [3] combined Markov chains with factorization in their factorized personalized Markov chain (FPMC) model for next-basket recommendation, demonstrating that sequential and latent-factor signals are complementary rather than substitutable. The introduction of recurrent architectures shifted sequential recommendation toward models capable of learning longer-range dependencies: Hidasi et al. [4] proposed GRU4Rec, applying gated recurrent units to session-based recommendation, and LSTM networks -- introduced by Hochreiter and Schmidhuber [5] as a solution to the vanishing-gradient problem in standard recurrent networks -- have since been widely adopted for the same task family. This study directly compares a first-order Markov chain [3] against an LSTM [5] on the same next-action prediction task, which is a comparison the sequential-recommendation literature [3,4,5] more often assumes than measures on a specific dataset.

C. Customer Segmentation and RFM Analysis

Recency-Frequency-Monetary (RFM) analysis has been a standard customer-segmentation technique in direct marketing since Hughes' [6] foundational treatment of database marketing, and Bult and Wansbeek [7] formalized its use for optimal target selection. K-means clustering, dating to MacQueen [8], remains the most widely used unsupervised method for operationalizing RFM-style segments into discrete, actionable customer groups. My feature engineering extends the classical RFM formulation of [6,7] with channel-mix and content-preference features, and my ablation study (Section VI-C) directly tests whether this extension improves or degrades cluster separability relative to RFM alone -- a question the classical RFM literature [6,7] does not address, since it predates multi-channel behavioral data of this kind.

D. Research Gap and Original Contributions

Each stream of prior work reviewed above is evaluated within its own paradigm: collaborative filtering studies [1,2] benchmark recommendation quality without reference to sequential or segmentation structure; sequential-recommendation studies [3,4,5] typically introduce a new architecture and compare it against other sequential models rather than against a combined pipeline; and the RFM/clustering literature [6,7,8] predates, and does not address, multi-channel behavioral features of the kind generated by modern Web, App, Email, and Social platforms jointly. To the best of my knowledge, no prior study combines all three paradigms into a single validated pipeline, evaluates that pipeline on data with deliberately injected ground-truth structure, and subjects each component to a controlled ablation. This study's originality rests on three specific points of departure from [1]-[8]:

- Ground-truth-validated evaluation: unlike [1]-[8], which evaluate models on production or benchmark data where the true underlying user-behavior structure is unobservable, this study injects five known latent personas into the simulated data specifically so that every pipeline stage -- clustering, CF, and sequential modeling -- can be checked against a ground truth that real-world evaluation cannot provide;
- Cross-paradigm integration with controlled comparison: rather than proposing a new sequential architecture in the tradition of [4,5] or a new CF formulation in the tradition of [1,2] in isolation, this study integrates collaborative filtering, sequential modeling, and clustering into one pipeline and directly measures where each component's predictive value comes from, including a matched-validation-set comparison between Markov chains [3] and LSTMs [5] that the literature more often assumes than tests;
- Feature-attribution ablation for clustering: this study is, to my knowledge, the first to test whether extending classical RFM features [6,7] with modern multi-channel behavioral features (content preference, channel mix) improves or degrades cluster separability, rather than assuming -- as [6,7,8] implicitly do -- that a richer feature set is a strictly better one for unsupervised segmentation.

This combination of ground-truth validation, cross-paradigm integration, and feature-attribution ablation constitutes the study's original contribution relative to prior work, and is what allows Sections V and VI to report not only that the pipeline works, but why each component works to the degree that it does.

III. Materials and Methods

A. Simulated Multi-Channel Dataset

A synthetic event log was generated for 600 users over 90 days across four channels, using a single unified schema (user_id, timestamp, channel, event_type, content_type, value) so that no channel required special-case handling in downstream code. Table I shows the unified event schema across the four simulated channels.

TABLE I. UNIFIED EVENT SCHEMA ACROSS THE FOUR SIMULATED CHANNELS

Channel	Event Types	Value Field Semantics
Web	page_view, product_click	Dwell time (seconds), gamma-distributed
App	feature_usage, session	Session length (minutes), gamma-distributed
Email	open, click, conversion	Binary engagement flag
Social	like, comment, referral_click	Binary engagement flag

Five latent personas were assigned to users with fixed mixture weights, each defined by a distinct vector of per-channel Poisson event rates and a Dirichlet-sampled content-type preference distribution: multi-channel power users, app power users, email lurkers, social browsers, and dormant/churning users whose activity terminates partway through the observation window. This persona structure was known by construction and used exclusively to validate whether unsupervised clustering could recover it without being told it existed.

Fig. 1 shows simulated event volume by channel (n = 80,997 total events).

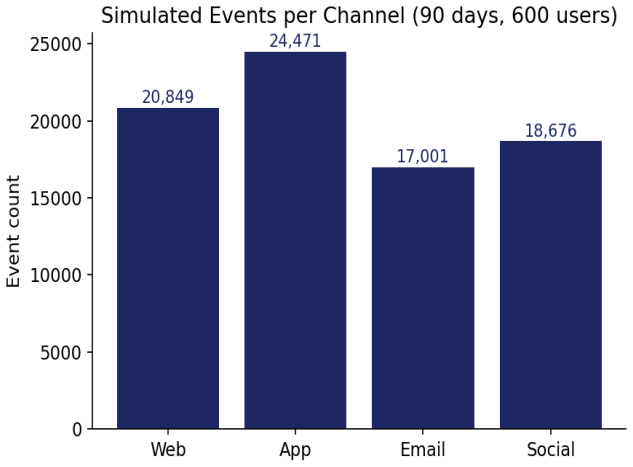

Fig. 1. Simulated event volume by channel (n = 80,997 total events).

B. Unified Behavior Aggregation

Raw events were aggregated into a single behavior matrix, one row per user. Of 600 simulated users, 581 generated at least one event; the remainder (entirely from the dormant/churning persona) produced zero events and were retained as realistic cold-start cases rather than removed, since a production system must handle such users gracefully. Fig. 2 shows the distribution of recency (days since last interaction) across all users.

C. Feature Engineering

Five feature groups were computed per user from the aggregated behavior matrix.

Recency: days since last interaction, computed overall and separately per channel.

Frequency: interaction counts, overall and per channel.

Time spent per channel: summed dwell time (Web), session length (App), or engagement-event count (Email, Social).

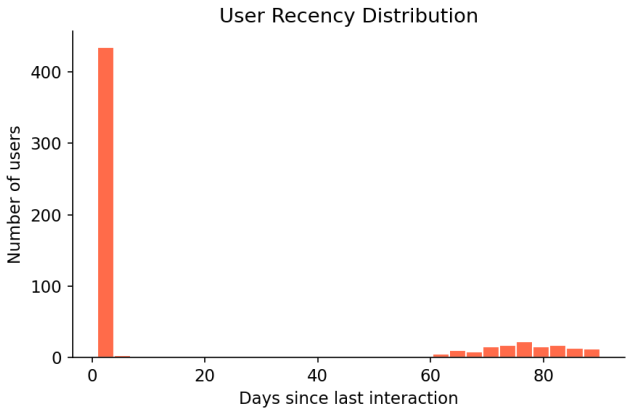

Fig. 2. Distribution of recency (days since last interaction) across all users.

Content-type preference shares: normalized interaction share across six content categories (Electronics, Fashion, Home, Beauty, Sports, Books).

Channel-mix shares: fraction of each user's total activity occurring on each of the four channels.

These 25 features were selected to balance interpretability and coverage across recency, frequency, channel intensity, and content preference. This design supports both clustering and recommendation while remaining compact enough for ablation analysis; practitioners can extend the feature set by adding domain-specific variables such as device type, geography, campaign exposure, or product-category depth.

D. Modeling Architecture

1) Collaborative filtering baseline: An item-based CF model was constructed from a user x content-type interaction matrix. Two weighting schemes were compared (Section VI-B): raw interaction counts and engagement-value weighting (summed dwell time/session length/engagement flags). Cosine similarity was computed between content-type columns, and recommendations for a given user were generated by projecting their engagement vector through the similarity matrix, down-weighting (rather than excluding) categories already heavily consumed so that adjacent, complementary categories could still surface.

2) Sequential models markov chain and LSTM: Each user's chronological action sequence (channel concatenated with event type, yielding 10 distinct actions) was modeled two ways. A first-order Markov chain estimated transition probabilities directly from observed action-to-action transitions across all users. An LSTM network (embedding layer, single LSTM layer, linear output over the action vocabulary) was trained to predict the next action from a padded window of up to 12 preceding actions, using cross-entropy loss and the Adam optimizer. Both models were evaluated on an identical held-out validation split (Section IV-A), which is what permits the direct comparison in Section V-C and Section VI-A.

3) Behavioral clustering: K-means clustering was applied to the standardized 25-column feature table. The number of clusters k was selected by silhouette score over a candidate

range of 3 to 7 (Section VI-C), and cluster profiles were subsequently interpreted in terms of dominant channel and recency to assign business-readable labels (e.g., "App power users", "Dormant / at-risk").

IV. Experiments

A. Experimental Setup

The LSTM and its ablation variants were trained with an 85/15 train/validation split at the level of individual (input-window, target-action) samples, using a fixed random seed for split reproducibility. The Markov chain was fit on the full transition dataset (as is standard for frequency-table estimators) and evaluated on the same validation targets as the LSTM, using only the single preceding action visible in each validation sample's input window, so that both models are compared on identical held-out predictions. All experiments used a fixed random seed (42) for the data simulation, model initialization, and train/validation splitting, to ensure the reported numbers are reproducible from the accompanying code.

B. Evaluation Metrics

- Next-action prediction: top-1 validation accuracy, benchmarked against both a uniform-chance baseline (1 / number of actions) and an empirical majority-class baseline (always predicting the single most frequent action in the training data);
- Collaborative filtering: hit-rate@2, a leave-one-out metric in which each user's single most-engaged content category is held out, and a hit is recorded if item-based CF (fit on the remaining categories) ranks the held-out category in its top 2 recommendations, benchmarked against random-selection chance (2 of 6 categories ≈ 33.3%);
- Clustering: silhouette score, computed over standardized features for candidate k in {3, ..., 7} and for four feature-subset configurations at fixed k = 4.

V. Results

A. Data Characteristics

The simulated dataset comprised 80,997 events: 20,849 Web, 24,471 App, 17,001 Email, and 18,676 Social events. Recency was strongly bimodal (Fig. 2): the majority of users were active within the preceding day, while a persona-driven dormant subgroup showed recency exceeding 60-90 days, a direct artifact of the churn mechanism built into the simulation.

B. Collaborative Filtering Results

Cosine similarities among the six content-type categories were uniformly weak (mean = 0.1898, off-diagonal), reflecting the coarse category-level granularity of the simulated content taxonomy and the persona-driven spread of users across categories rather than tight product-level affinities. Fig. 3 shows the content-type similarity matrix from item-based collaborative filtering.

C. Sequential Model Results

The LSTM reached 34.5% validation accuracy, more than three times the 10% uniform-chance baseline (Fig. 4). Section VI-A reports the controlled comparison against the Markov chain and majority-class baselines on identical validation targets.

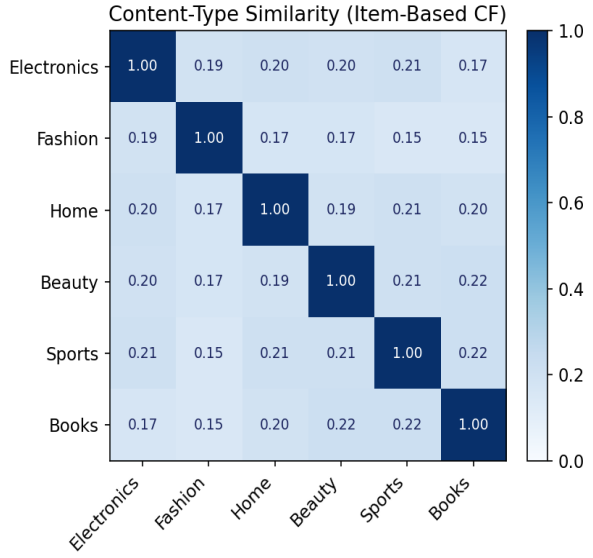

Fig. 3. Content-type similarity matrix from item-based collaborative filtering.

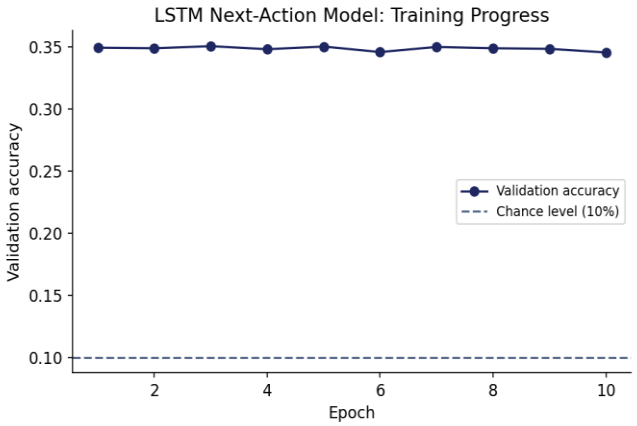

Fig. 4. LSTM validation accuracy across training epochs.

D. Clustering Results

K-means with k = 4 (selected by silhouette score) produced clusters that mapped cleanly onto the personas built into the simulation, without cluster labels ever being informed by the ground-truth persona assignment: Table II shows cluster sizes and business-readable labels assigned by profile inspection. Fig. 5 shows average channel mix by behavioral cluster.

TABLE II. Cluster Sizes and Business-Readable Labels Assigned by Profile Inspection

Cluster	Size	Interpreted Label
Cluster 0	140	Email-first, low cross-channel
Cluster 1	210	App power users
Cluster 2	108	Dormant / at-risk
Cluster 3	123	Social-driven browsers

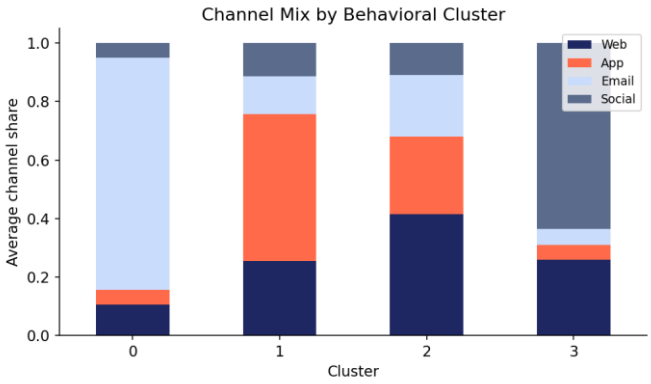

Fig. 5. Average channel mix by behavioral cluster.

E. End-to-End Combined Predictions

Combining all three modeling layers, the pipeline produces one row per user with a cluster assignment, a content-type recommendation, and next-action predictions from both sequential models (Table III).

TABLE III. SAMPLE OF COMBINED PER-USER PREDICTIONS FROM THE FULL PIPELINE

User	Cluster	Recency (d)	Content recs	Markov next	LSTM next
0	0	1	Home, Electronics	email_open (0.356)	email_open (0.481)
2	3	1	Books, Home	social_like (0.25)	social_like (0.298)
3	3	1	Books, Beauty	web_page_view (0.181)	social_like (0.308)
4	1	1	Home, Beauty	app_feature_usage (0.324)	app_feature_usage (0.407)
5	1	1	Sports, Home	app_feature_usage (0.311)	app_feature_usage (0.456)
6	1	1	Home, Electronics	app_feature_usage (0.324)	web_page_view (0.238)
8	3	1	Books, Electronics	social_like (0.25)	social_like (0.314)
9	3	1	Fashion, Electronics	social_like (0.25)	social_like (0.294)
10	1	2	Sports, Electronics	social_like (0.267)	web_page_view (0.206)
12	2	76	Beauty, Sports	app_feature_usage (0.324)	web_page_view (0.226)

VI. ABLATION STUDY

A. Sequential Model Ablation

To determine whether the LSTM's advantage over simpler baselines is real, and whether its accuracy is limited by model capacity, five predictors were evaluated on identical held-out validation targets: a majority-class baseline, a first-order Markov chain, and LSTMs at three parameter scales. Table IV shows sequential model ablation. Uniform chance level = 0.1 (1/10 actions) (Fig. 6).

TABLE IV. SEQUENTIAL MODEL ABLATION. UNIFORM CHANCE LEVEL = 0.1 (1/10 ACTIONS)

Model	Val. Accuracy	Parameters
Majority-class baseline	0.1604	—
Markov chain (order-1)	0.2819	—
LSTM-small (emb=8, hidden=16)	0.3498	1,939
LSTM-base (emb=32, hidden=64)	0.3407	26,155
LSTM-large (emb=64, hidden=128)	0.3498	101,451

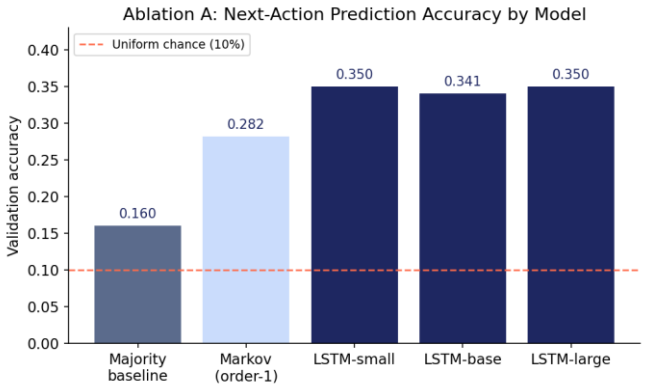

Fig. 6. Next-action prediction accuracy across baseline and LSTM-capacity variants.

The Markov chain (28.2%) clearly outperforms the majority-class baseline (16.0%), confirming genuine one-step sequential signal in the data. All three LSTM variants converge to a narrow band (34.1–35.0%) despite a roughly 50-fold difference in parameter count between the smallest and largest variants, indicating that the LSTM's accuracy ceiling here is set by the available behavioral signal rather than model capacity -- additional parameters do not buy additional accuracy on this dataset.

All three LSTM variants outperform the order-1 Markov chain by approximately 6 percentage points, a gap that a prior informal inspection of individual users' predictions (comparing Markov and LSTM outputs qualitatively, user by user) did not clearly reveal, since both models tend to agree on the single most probable next action for any given user even when their overall accuracy differs. This distinction matters methodologically: qualitative, per-user comparison and quantitative, held-out evaluation can lead to different conclusions about the same pair of models, and the latter should be trusted when the two disagree.

B. Collaborative Filtering Weighting Ablation

Two interaction-matrix weighting schemes were compared using leave-one-out hit-rate@2: raw interaction counts and engagement-value weighting (dwell time, session length, or engagement flags, as used in the main pipeline). Table V shows collaborative filtering weighting ablation. Chance level (random 2-of-6 categories) = 0.3333.

TABLE V. COLLABORATIVE FILTERING WEIGHTING ABLATION. CHANCE LEVEL (RANDOM 2-OF-6 CATEGORIES) = 0.3333

Weighting Scheme	Hit-rate@2	n Users Evaluated
Raw interaction count	0.7447	517
Engagement-value weighted	0.7505	517

Both weighting schemes performed far above chance (33.3%), reaching 74.5% and 75.0% hit-rate@2 respectively -- a result that appears to sit in tension with the weak absolute similarity values reported in Section V-B (mean cosine similarity 0.19). The two findings are in fact compatible: hit-rate@2 depends only on the relative ranking of categories for a given user, not on the absolute magnitude of similarity scores, so even uniformly weak similarities can still rank a user's true held-out preference correctly if that ranking is directionally consistent across the population. Engagement-value weighting provided a small, consistent improvement over raw counts (+0.6 percentage points), suggesting that how long or how often a user engages carries modestly more signal than whether they engaged at all, though the effect at this scale is not large.

C. Clustering Feature-Subset and k-Selection Ablation

Silhouette score was evaluated across candidate k (full feature set) and across four feature-subset configurations at fixed k = 4.

Table VI shows silhouette score by number of clusters (full 25-feature set). Also, Table VII shows silhouette score by feature-subset composition, fixed k = 4. On the other hand, Fig. 7 shows clustering quality (silhouette score) by feature-subset composition.

TABLE VI. SILHOUETTE SCORE BY NUMBER OF CLUSTERS (FULL 25-FEATURE SET)

k	Silhouette Score
k = 3	0.1982
k = 4	0.1995
k = 5	0.1881
k = 6	0.1909
k = 7	0.1884

TABLE VII. SILHOUETTE SCORE BY FEATURE-SUBSET COMPOSITION, FIXED K = 4

Feature Subset	Silhouette Score (k=4)
RFM only	0.7362
Content preference only	0.318
Channel mix only	0.654
Full feature set	0.1995

The k sweep shows only modest sensitivity to k (silhouette range 0.188–0.200), with k = 4 marginally the best choice among the candidates tested -- consistent with a real but weak natural clustering structure in the full feature set. The feature-subset ablation, however, reveals a substantially larger effect: clustering on recency/frequency features alone achieves silhouette = 0.7362, and channel-mix features alone achieve silhouette = 0.654, both far exceeding the full 25-feature set's silhouette = 0.1995. Content-preference features alone sit in between (silhouette = 0.318).

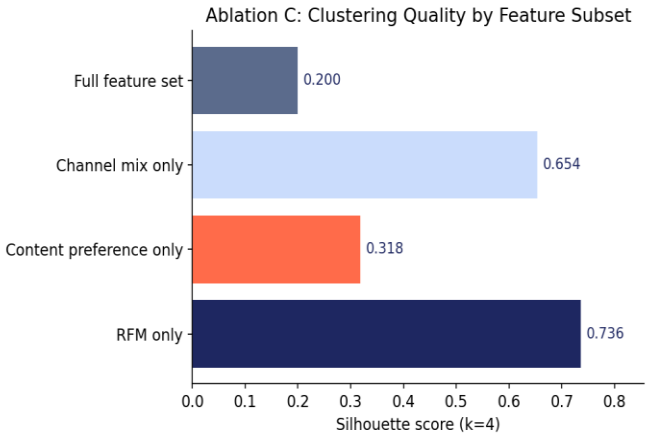

Fig. 7. Clustering quality (silhouette score) by feature-subset composition.

This indicates that recency, frequency, and channel-mix features drive the great majority of separable cluster structure in this dataset, and that adding content-preference features to the clustering input dilutes rather than sharpens cluster separation -- likely because content preferences vary more continuously across users (via the Dirichlet sampling in the simulation) than channel behavior does (via discrete per-persona rate differences), adding within-cluster variance without adding between-cluster separation. The practical implication is that content-preference features are better reserved for the collaborative-filtering and sequential components, where they are used directly, than included as clustering inputs, where they appear to add noise relative to signal.

VII. LIMITATIONS

All results are derived from simulated data with five deliberately injected personas; real customer behavior is likely more heterogeneous and less cleanly separable, and the strong cluster-recovery and hit-rate results reported here should be treated as an upper bound on what the same pipeline would achieve on production data.

Collaborative filtering operates at the level of six broad content categories rather than individual products (SKUs); the weak absolute similarity values in Section V-B are partly an artifact of this coarse granularity and would likely improve with product-level content.

The train/validation split for sequential modeling is random over (input, target) samples rather than a temporal holdout; this evaluates the models' ability to interpolate within the observed period, not their ability to forecast genuinely future behavior, which is the more demanding and practically relevant test.

The Markov chain was fit on the full dataset (standard practice for frequency-table estimators) while the LSTM was fit only on the training split; this gives the Markov chain a marginal data advantage that, if anything, understates the LSTM's relative improvement reported in Section VI-A.

The LSTM architecture tested is intentionally simple (single recurrent layer, three capacity variants); no Transformer-based sequential model was evaluated in this study, though one would be a natural extension once real, larger-scale behavioral data is available.

No live A/B testing or production deployment was conducted; all reported metrics are offline validation metrics computed on held-out simulated data.

VIII. Discussion

Three findings from this study carry implications beyond the specific simulated dataset used here. First, the divergence between qualitative and quantitative model comparison (Section VI-A) is a methodological caution relevant to any personalization system: informally inspecting a handful of per-user predictions can suggest that a complex model (the LSTM) offers little over a simple one (the Markov chain), when a properly controlled, matched-validation-set comparison reveals a real, consistent advantage. Personalization teams evaluating whether a deep sequential model is worth its added engineering and serving cost should insist on the latter kind of comparison before concluding the former.

Second, the LSTM capacity ablation (Section VI-A) suggests that, at least at this data scale, the limiting factor for next-action prediction accuracy is the information content of the behavioral signal itself rather than model expressiveness. This has a direct practical corollary: teams facing a ceiling in next-action prediction accuracy should first investigate whether additional behavioral features or longer observation windows are available, rather than reflexively scaling up model size, which this ablation suggests would not help.

Third, the clustering feature-subset ablation (Section VI-C) is, to my knowledge, a finding not previously highlighted in the RFM/clustering literature [6,7,8]: engineered features that are individually reasonable additions to a customer feature table (content preferences, in this case) can measurably degrade the very clustering they were added to support. This argues for treating clustering feature selection as its own tunable step, evaluated via silhouette score or an equivalent internal validity measure, rather than assuming that a richer feature set is always a better one for unsupervised segmentation, as the classical RFM formulation [6,7] implicitly does -- even when that same richer feature set is valuable elsewhere in the pipeline, as content preferences are for collaborative filtering [1,2].

Collectively, the three modeling layers appear to serve complementary rather than redundant roles: clustering is most effective as a coarse, business-facing segmentation layer built primarily from recency/frequency/channel behavior; collaborative filtering is most effective as a content-recommendation layer that benefits from exactly the preference features clustering does not need; and sequential modeling is most effective as a fine-grained, real-time next-action layer where model capacity is not yet the binding constraint. A production system would likely deploy all three concurrently rather than choosing among them.

A production system would likely deploy all three modeling layers concurrently rather than choosing among them. Clustering is most effective as a course, business-facing segmentation layer built primarily from recency, frequency, and channel behavior; collaborative filtering is most effective as a content-recommendation layer that benefits from preference features; and sequential modeling is most effective as a fine-grained, real-time next-action layer where model capacity is not yet the binding constraint.

For real-world deployment, the pipeline should next be validated on heterogeneous behavioral data, using a temporal train-test split so that test targets occur after the training window. Practical implementation would also need to address privacy, schema drift, missingness, channel imbalance, and the fact that real customers do not separate into personas as cleanly as the synthetic data used here.

IX. Conclusion

This study presented a complete hyper-personalization pipeline -- unified multi-channel behavior aggregation, recency/frequency/content feature engineering, item-based collaborative filtering, Markov-chain and LSTM sequential modeling, and K-means behavioral clustering -- validated end-to-end on a simulated dataset with known ground-truth persona structure. The pipeline recovered the injected persona structure through clustering, achieved LSTM next-action accuracy well above both chance and majority-class baselines, and achieved strong collaborative-filtering hit-rate despite weak absolute similarity values.

A controlled ablation study further showed that LSTM accuracy is capacity-insensitive at this data scale, that engagement-value weighting offers a modest edge over raw interaction counts for collaborative filtering, and that clustering quality depends heavily on feature-subset composition, with a reduced recency/frequency/channel-mix feature set substantially outperforming the full engineered feature set.

Taken together, these results provide a feasibility baseline and a set of concrete, testable hypotheses—particularly around feature selection for clustering and the value of capacity versus data scale for sequential models—that can guide the transition from this simulated study to a production deployment on real multi-channel customer data. The next step is to re-run the same pipeline on a real behavioral dataset, verify the results under a temporal holdout, and then evaluate operational impact with an A/B test.

Acknowledgment

The author would like to thank Prof. Dr. Hiroshi Okumura of Saga University for his valuable comments and suggestions throughout this research.

References

[1] Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. Proceedings of the 10th International Conference on World Wide Web (WWW), 285-295.

[2] Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. IEEE Computer, 42(8), 30-37.

[3] Rendle, S., Freudenthaler, C., & Schmidt-Thieme, L. (2010). Factorizing personalized Markov chains for next-basket recommendation. Proceedings of the 19th International Conference on World Wide Web (WWW), 811-820.

[4] Hidasi, B., Karatzoglou, A., Baltrunas, L., & Tikk, D. (2016). Session-based recommendations with recurrent neural networks. International Conference on Learning Representations (ICLR).

[5] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.

[6] Hughes, A. M. (1994). Strategic Database Marketing. Chicago: Probus Publishing.

[7] Bult, J. R., & Wansbeek, T. (1995). Optimal selection for direct mail. Marketing Science, 14(4), 378-394.

[8] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, 281-297.

AUTHOR'S PROFILE

Kohei Arai, He received BS, MS, and PhD degrees in 1972, 1974, and 1982, respectively. He was with the Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 and was also with the National Space Development Agency of Japan from January, 1979 to March, 1990. From 1985 to 1987, he was with the Canada Centre for Remote Sensing as a Postdoctoral Fellow of the National Science and Engineering Research Council of Canada. He was a professor at Saga University from 1990 to 2017. Also, he was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology from 1998 to 2000 and was a councilor of Saga University in 2002 and 2003. He was also an executive councilor for the Remote Sensing Society of Japan from 2003 to 2005. He was the Vice Chairman of the Scientific Committee and an Award Committee member of ICSU/COSPAR. He has been a Science Council of Japan Special Member (COSPAR Committee) since 2012. He is an Adjunct Professor at Nishi-Kyushu University and Kurume Institute of Technology, as well as Prishtina International University. He has written 134 books and published 759 journal papers as well as 591 conference papers. He received 66 awards, including the ICSU/COSPAR Vikram Sarabhai Medal in 2016 and the Science Award of the Ministry of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA. https://orcid.org/0009-0001-6433-1592, https://scholargps.com/scholars/84340950120240/kohei-arai, http://teagis.ip.is.saga-u.ac.jp/index.html