From Clicks to Intent: Cross-Platform Session Embeddings with LLM-Distilled Taxonomy for Financial Services Recommendations

Dianjing Fan*
Capital One
McLean, VA, USA

Yao Li*
Capital One
New York, NY, USA

Kyaw Hpone Myint
Capital One
Boston, MA, USA

Dwipam Katariya
Capital One
McLean, VA, USA

Alexandre Day
Capital One
McLean, VA, USA

Pranab Mohanty
Capital One
McLean, VA, USA

Giri Iyengar
Capital One
McLean, VA, USA

Abstract

Sequential user behavior modeling is widely adopted in industrial recommender systems; however, significant gaps remain in financial services, where pre-login web interactions and authenticated in-app experiences differ drastically. Specifically, pre-login web users typically explore new products, whereas logged-in app users focus on account servicing. Due to the challenge of cross-channel entity resolution (e.g., matching anonymous web sessions to authenticated mobile accounts), web-based intent signals remain underutilized for post-authentication personalization. Existing methods for capturing web-based intent are often ad-hoc and narrow, lacking the flexibility to support both quantitative downstream recommendations and qualitative understanding at scale. In this work, we propose a scalable and dual-purpose intent prediction framework for web-based interactions and demonstrate its applicability for personalization. Our approach transforms raw web clickstreams into two outputs: a self-supervised Transformer encodes multi-modal clickstreams into a compact session embedding, while an LLM-based taxonomy generation and distillation pipeline produces interpretable intent labels. Our system demonstrates that self-supervised clickstream representations combined with LLM-distilled taxonomies can jointly serve quantitative tasks and qualitative understanding in production: on the mobile homepage tile ranking task, the session embedding improves macro Recall@1 by 1.88% and reduces Log Loss by 13.38% over production baselines. On the user conversion prediction task, the embedding outperforms the LLM labels by 4.3% on micro F1, while the distillation layer delivers interpretable labels at ultra-low latency with only a 7% performance drop.

CCS Concepts

• **Computing methodologies** → **Learning latent representations**; *Neural networks*; *Information extraction*; • **Information systems** → **Recommender systems**.

Keywords

transformers, sequential recommendation, intent taxonomy, tabular clickstream data, knowledge distillation, large language models

*Both authors contributed equally to this research.

ACM Reference Format:
Dianjing Fan, Yao Li, Kyaw Hpone Myint, Dwipam Katariya, Alexandre Day, Pranab Mohanty, and Giri Iyengar. 2026. From Clicks to Intent: Cross-Platform Session Embeddings with LLM-Distilled Taxonomy for Financial Services Recommendations. In *Proceedings of* . ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Large-scale financial services companies serve their users through digital platforms, including web and mobile applications, generating massive volumes of clickstream data, such as timestamps, page views, and user actions. These behavioral logs capture *what* users did, yet not *why*: the underlying intent remains latent. Moving beyond surface-level interactions to understand user intent would enable recommender systems to deliver more relevant personalized experiences [6, 13, 14]. In financial services, the user intent in the pre-login setting carries significant weight: unlike e-commerce or social platforms, where the richest behavioral signal is generated post-login, most financial product exploration, comparison, and even application initiation occurs on the public website before authentication, making pre-login browsing one of the highest-leverage signals available. However, two challenges prevent this signal from being used:

- Pre-login context is lost at login: recommender systems treat users as if they have no prior unauthenticated browsing history, discarding the intent they have already expressed on the website. This can happen because identifying users pre-login requires entity resolution methods and connecting this information post-login requires cross-platform event capture to be centralized.
- Existing intent representations tend to be ad-hoc and overly narrow (e.g., simply logging that "a user visited credit card pages"), thereby failing to capture more nuanced behavioral signals. Due to both the regulatory nature of the financial industry and the analytical needs to provide a better user experience, qualitative analysis and explainability are crucial. However, existing approaches lack an explainability-aware intent representation that serves both quantitative tasks (e.g.,

ranking, recommendation) and these essential qualitative use cases (e.g., segmentation, analytics) at scale.

A semantic intent engine that transforms raw clickstreams into structured intent representations would address these challenges, improving dynamic user segmentation, personalized content ranking, marketing arbitration, and much more. To connect pre-login browsing with post-login experiences, we leverage two straightforward identity signals: within-session linkage via the first-party cookie[1], and cross-device linkage from web to mobile app via the shared login credential. Building on this linkage to address the gap of discarded pre-login signals in financial services, we present a solution with three core contributions that demonstrate the practical usefulness of our approach:

- **Dual-purpose intent prediction system.** A self-supervised Transformer encodes multi-modal web clickstream into a session embedding, while an LLM-generated intent taxonomy is distilled into a lightweight classifier to provide interpretable intent labels from the same embedding space. This yields one shared representation with two complementary outputs: a dense embedding for quantitative tasks and human-readable labels for qualitative understanding.
- **Clustering-enhanced LLM taxonomy generation.** A clustering layer over session embeddings enables stratified sampling, improving the iterative taxonomy generation approach [21], which generates and refines an intent taxonomy through successive minibatches of samples, by ensuring the LLM sees the full diversity of user behavior rather than a random subset.
- **Embedding-space intent distillation.** LLM-generated intent labels are distilled into a lightweight classifier that operates directly on the compact fixed-dimensional vector as input rather than raw text sequences, allowing it to scale to 10M+ daily sessions at inference without LLM calls or separate text encoding.

2 Related Work

2.1 User Behavior Representation Learning and Sequential Recommendation

Modeling user behavior sequences is central to modern recommender systems, and methods differ primarily in how sequential dependencies are captured.

Early approaches apply traditional machine learning techniques to study user behavior, such as manual rule mining [12], Markov Chain models [1], and unsupervised clustering [15, 22]. These methods capture only static or short-range patterns and lack the capacity to model long behavior sequences. Recurrent architectures such as GRU4Rec [8] advance beyond these with gated sequential encoders, but are inherently non-parallelizable and suffer from decaying sensitivity to long-range dependencies.

[1]All clickstream data used in this study is collected from a single first-party domain and is handled under internal data governance policies.

Transformer-based architectures address these limitations with self-attention over behavior sequences. SASRec [10] applies unidirectional self-attention to capture both short- and long-term sequential dependencies. BERT4Rec [20] extends this with bidirectional attention and masked-item prediction to leverage context from both directions. Industry systems scale these to multi-feature settings: BST [5] at Alibaba and TransAct [26] at Pinterest enrich item sequences with action-level signals; TRACE [2], STEP [19], and TIMeSynC [11] extend further to tabular event streams where each event is a heterogeneous row of categorical and numerical attributes. However, these systems fuse raw categorical and numerical features without incorporating enriched semantic representations, such as pretrained text embeddings derived from page content. Additionally, none provides an interpretability layer over the modeled behavior sequences: the resulting representations cannot be inspected, summarized, or grounded in human-readable form.

2.2 LLM-Based Taxonomy Generation

In recommender systems, intent representations have been used to structure candidate generation [29], model multiple user interests [13], and encode latent intents in sequential models [6]. While these methods encode intent implicitly, an explicit, human-readable taxonomy that organizes intents into structured categories would be a valuable complement when goals extend to behavioral analysis, user segmentation, and directly guiding LLM-based ranking [14].

Pre-LLM taxonomy generation relied on hand-crafted patterns [4], hierarchical topic modeling [3], and embedding-based term clustering [27], but either lacked scalability or failed to capture nuanced semantics. Large language models bridge this gap by producing fluent, interpretable labels over groups of documents, as demonstrated by methods such as TopicGPT [16], GoalEx [24], and ClusterLLM [28]. TnT-LLM [21] shows that iterative minibatch refinement paired with human-in-the-loop validation [18] is needed for large datasets that do not fit into a single prompt. To make LLM-derived taxonomies deployable at production scale, these systems distill LLM-generated labels into lightweight student classifiers [21, 25], consistent with broader evidence that LLM teachers can train compact models at a fraction of the inference cost [9, 23]. However, all prior LLM-based taxonomy systems operate on raw text inputs. Applying iterative LLM-based taxonomy generation and distillation based on the intent embeddings derived from multi-modal clickstream sequences in financial services remains unexplored.

In summary, to our knowledge, prior work addresses user behavioral representation learning (Section 2.1) and LLM-based taxonomy generation and distillation separately, but none combines the two.

3 Method

3.1 Overview

Our system produces two complementary outputs from pre-login web clickstream data: (1) **a dense session embedding** for downstream tasks, and (2) **human-readable intent labels** for interpretability.

In the first stage (Section 3.2), raw clickstream events are encoded into session embeddings via a self-supervised Transformer. In the second stage (Section 3.3), an LLM-driven pipeline constructs an

intent taxonomy from representative sessions and assigns each session to multiple categories. The resulting labels are distilled into a lightweight classifier that consumes the embeddings and produces interpretable intent labels for scalable inference.

3.2 Multi-Modal Session Embedding

The input to our model is pre-login web clickstream sessions: chronologically ordered sequences of events. Let a session be $s = (x_1, \ldots, x_T)$ of length T (the maximum session length, with shorter sessions padded), where each event x_i carries f heterogeneous features, including page information, dwell time, event type (click, scroll, view, etc.), and semantically enriched features such as page HTML content. All features are encoded into a common d-dimensional space.

Discrete features via learned embedding tables. Each discrete feature is embedded via a learned lookup table. A feature with vocabulary size V has a learned embedding matrix $E \in \mathbb{R}^{V \times d}$, and its token index is mapped to a d-dimensional vector by row lookup.

Page content via a pretrained sentence encoder. To inject semantic content from the rendered page rather than just the URL string, we pre-compute a vector for every URL by fetching its HTML, extracting the main text, and encoding it with a pretrained sentence transformer [17] into a d_{url}-dimensional vector. The resulting vector is then projected to d dimensions through a learned linear layer $W_{\text{url}} \in \mathbb{R}^{d \times d_{\text{url}}}$ with bias $b_{\text{url}} \in \mathbb{R}^d$.

Sequence position via a learned positional embedding. A learned positional embedding matrix $P \in \mathbb{R}^{T \times d}$ supplies one d-dimensional vector per position, where the i-th row of P is the embedding for position $i \in \{1, \ldots, T\}$.

Per event, the feature vectors are concatenated into an fd-dimensional vector and fused by an $\text{MLP} : \mathbb{R}^{fd} \to \mathbb{R}^d$, after which the positional embedding is added and a Transformer encoder $\mathcal{T}$ attends over the full sequence:

$$h_i = \text{MLP}\big(e_i^{(1)} \oplus \cdots \oplus e_i^{(f)}\big) \in \mathbb{R}^d, \tag{1}$$

$$z_i = h_i + P_i \in \mathbb{R}^d, \tag{2}$$

$$(o_1, \ldots, o_T) = \mathcal{T}(z_1, \ldots, z_T), \quad o_i \in \mathbb{R}^d, \tag{3}$$

where $\oplus$ denotes vector concatenation. Because the complete pre-login session is available at inference time rather than streamed incrementally, the encoder $\mathcal{T}$ is free to use either unidirectional or bidirectional self-attention, and the final session embedding is obtained by either last-hidden-state or mean pooling over $(o_1, \ldots, o_T)$. The model is trained self-supervised: depending on the attention setting, it predicts the page title and URL of either next event (causal) or masked events (bidirectional) from context. This requires no labeled data and encourages the embedding to capture the semantic content and sequential structure of the session.

Depending on the model and downstream performance (see Section 4.2), the final model is a multi-layer Transformer encoder trained with bidirectional masked language modeling and mean pooling.

3.3 Intent Taxonomy Generation and Distillation

While the session embeddings capture user behavior in a dense vector, they lack interpretability. To bridge this gap, we construct an intent taxonomy using LLMs and distill it into a lightweight classifier that maps embeddings to human-readable intent labels. The pipeline has three phases: clustering-based sampling, taxonomy generation, and label assignment and distillation.

Clustering-based sampling. We compare K-Means and HDBSCAN for clustering the session embeddings, tuning K for K-Means and minimum cluster size for HDBSCAN via silhouette score, Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI). We adopt K-Means with an optimal K based on these metrics. We then sample N sessions per cluster, yielding $K \times N$ representative sessions. This ensures the LLM sees the full diversity of user behavior rather than a random or skewed subset. Note that clustering serves only as a stratified sampling mechanism: clusters do not become taxonomy categories since individual sessions, particularly those farther from cluster centroids, may carry multiple intents. Each sampled session's raw clickstream is formatted as a sequence of `[dwell_time] [event_type] page_title : page_url` entries for LLM consumption.

Taxonomy generation. Because page metadata changes frequently as content is produced dynamically, manually curating and maintaining an intent taxonomy is impractical; an automated generation approach is needed to keep pace with evolving content while reducing manual overhead. Inspired by TnT-LLM [21], we adopt the iterative minibatch approach to construct a fixed intent taxonomy through three stages: *generate*, *update*, and *review*. Unlike free-form labeling, where an LLM may produce semantically equivalent but inconsistently worded intents, this approach constrains the label set and refines it iteratively as more data is observed.

In the first step, a *generation prompt* prompts the LLM to take the first minibatch of sessions and produce an initial taxonomy of I intent categories, each with a name, description, and confidence score. In subsequent steps, an *update prompt* asks the LLM to refine the taxonomy based on the new minibatches and the confidence scores of existing categories by adding, removing, merging, or splitting categories, while maintaining the target category count. A final *review prompt* checks for completeness, mutual exclusivity, and clarity. The generate–update–review cycle is repeated for E epochs over the sampled sessions. With $K \times N$ sampled sessions, a minibatch size of B, and E epochs, this requires $O(K \cdot N \cdot E/B)$ LLM calls. Each prompt includes background context, task instructions, formatting constraints (e.g., target number of categories, maximum intent label length), along with a minibatch of formatted clickstream sessions. After each taxonomy is produced, we perform a manual human review and refine the prompts as needed, ensuring the final taxonomy reflects domain expertise rather than LLM output alone.

Label assignment and distillation. With the taxonomy fixed, we assign intent labels to each sampled session via zero-shot LLM prompting and then train a distillation model. Following findings that LLMs achieve strong performance on annotation tasks [7], we use LLMs to classify each session into multiple intents from the taxonomy, each with a confidence score. Micro intents (specific subgoals) and chain-of-thought reasoning are additionally collected to

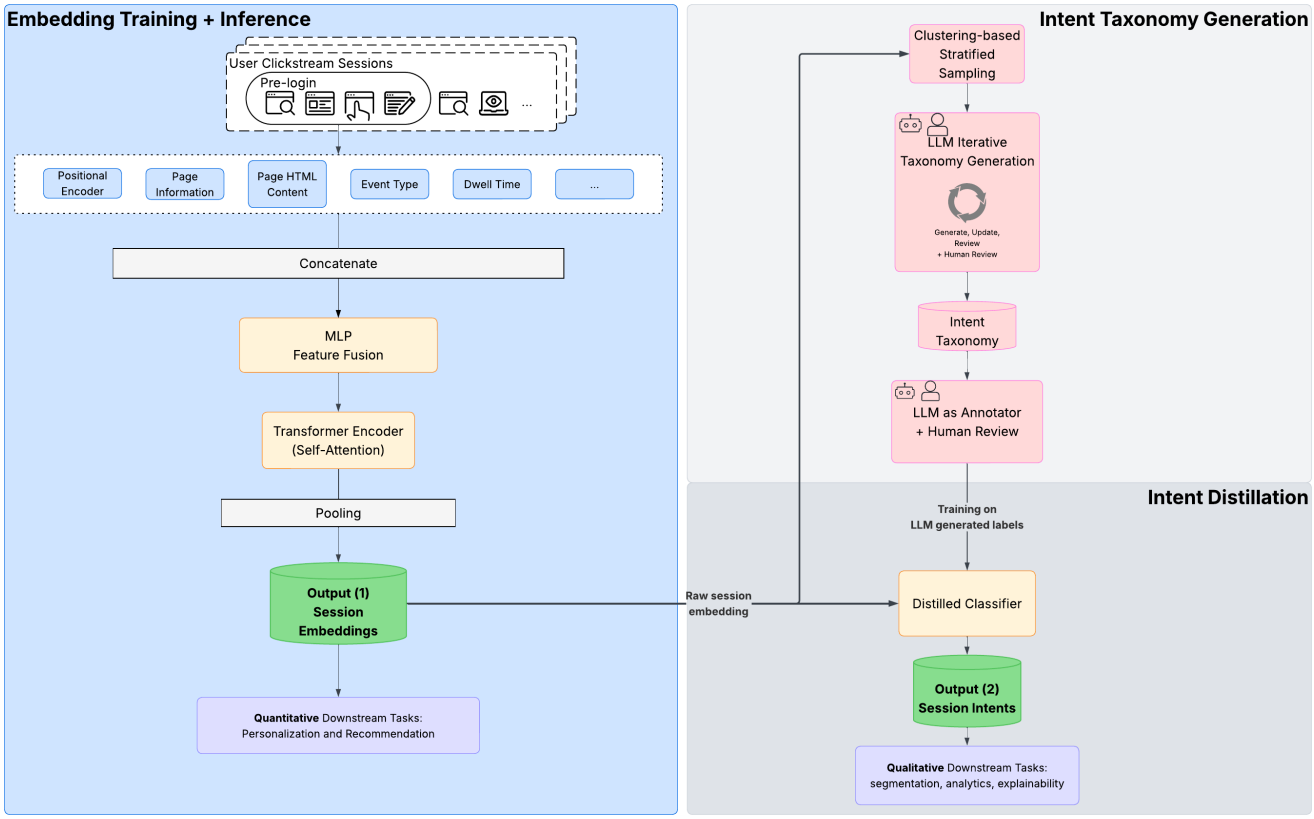

Figure 1: System architecture overview. Left (Section 3.2): multi-modal clickstream events are fused and encoded into session embeddings. Right (Section 3.3): clustering-based sampling, iterative taxonomy generation, label assignment, and distillation.

enable human review of the assigned labels, iteratively refining the prompt as needed.

To enable scalable inference, we distill the LLM-generated labels into a lightweight MLP classifier [21, 25]. The student model takes the session embedding as input and predicts intent probabilities independently for each label. We use the LLM's confidence scores as soft targets via per-label binary cross-entropy, preserving the teacher's uncertainty and producing better-calibrated predictions. At inference time, the distilled MLP is able to produce intent labels at scale without separate text encoding or LLM calls.

4 Experiments

4.1 Experimental Setup

Web Clickstream Data. Our primary data source is internal clickstream data from our company's financial services website. All activity occurs within a single first-party domain with no cross-site tracking involved. Each session consists of a chronologically ordered sequence of user interaction events, beginning with the user's first page visit and ending with idle timeout or logout; users may or may not log in during a session. Each event carries timestamps, page information, event type, cookies, device metadata, etc. For our study, we select the pre-login portion of each session that captures user intent before authentication. The full website receives approximately 10M sessions per day. We filter to sessions containing both pre-login activity (e.g., visits to product pages and servicing information) and post-login activity (e.g., account management, payments, product enrollment), yielding approximately 100K sessions per day.

To evaluate the session embeddings and intents on downstream tasks, we join pre-login sessions with post-login signals (mobile homepage tile clicks and product conversions) using a temporal join to link pre-login browsing activity to authenticated outcomes. We select four months of data for our experiments. All train, validation, and test splits are partitioned by time to prevent data leakage.

Mobile Homepage Tile Ranking. The internal mobile home screen presents various product and content tiles (Table 1) after users log in. Each tile is powered by its own recommender model that produces its content, while a global ranking model orders the tiles on the screen. This downstream validation dataset records which tiles each user taps, providing a direct signal of cross-platform post-login intent. We frame this as a multi-label classification task: given a user's features $\mathbf{x}$, predict which tiles the user will tap, i.e., $P(\text{click on tile } t \mid \mathbf{x})$ for each tile t in the candidate set. The production model is a multi-label classifier that ranks tiles using features derived from post-login behavior (e.g., sequences of page views) and

Table 1: Example home screen tile categories.

Tile Category	Example
Product	Internal product promotions
Travel	Flight and hotel deals
Shopping	External vendor shopping offers
Referral	Referral rewards

Table 2: Example user conversions.

Conversion
Open credit card
Apply for auto loan
Enroll in paperless statements
Book with travel portal

user context, with click-through rates per tile as a simple baseline. None of the existing features capture pre-login signals.

User Conversion. The other validation dataset focuses on user conversions across both mobile and web platforms, specifically tracking which products or features users ultimately apply for or enroll in after pre-login browsing (Table 2). We frame this as a multi-label classification task: given the pre-login intent signals, predict which post-login actions a user will take; multi-label classification is appropriate because a single session may lead to multiple conversions. This task is motivated by a key constraint in financial services: for unauthenticated users, especially prospects who have no prior relationship with the institution, pre-login clickstream activity is the *only* available signal, as static user features and post-login behavioral data do not yet exist. We therefore evaluate how well session embeddings and distilled intent labels perform in this data-scarce setting.

4.2 Mobile Homepage Tile Ranking Results

We evaluate three session embedding variants, each added as an additional feature set on top of a baseline model:

(1) Causal attention with last-hidden-state pooling;
(2) Causal attention with mean pooling;
(3) Bidirectional attention with mean pooling.

The baseline model is a LightGBM classifier trained one-vs-rest over tile categories, using click-through rate (CTR) and site retargeting features (SRF) as inputs. Both site retargeting features and session embeddings are derived from pre-login browsing activities: the former provides coarse-grained intent signals by aggregating page visit counts by product category over rolling time windows (9 dimensions), while the latter encode sequential patterns and content-level semantics. The evaluation metrics are macro Recall@K, which measures the fraction of users whose clicked tile appears in the top K predictions, and Log Loss, which measures the calibration of predicted probabilities. We report macro rather than micro metrics because of the heavy class imbalance in tile clicks: the most-clicked tile is a low-value servicing widget, while less-frequent tiles such as product offers and travel rewards are clicked less but drive more significant revenue.

Table 3: Overall macro Recall@K and Log Loss on the home screen tile ranking task. Values are percentage changes relative to baseline; for Log Loss, negative values indicate improvement.

Embedding Variant	R@1	R@2	R@3	Log Loss
CTR + SRF (baseline)	—	—	—	—
Causal + Last Hidden	+0.56%	+0.20%	+0.70%	+1.37%
Causal + Mean Pool.	+1.50%	+1.20%	+1.32%	−9.69%
Bidir. + Mean Pool.	**+1.88%**	**+1.16%**	**+1.29%**	**−13.38%**

Table 4: Example intents generated by the taxonomy pipeline at two granularity levels. I=25 serves as a compact catalog; I=35 adds finer-grained intents.

Category	Intent	I
Credit Cards	Credit Card Preapproval	25
	Credit Card Benefits	35
Bank Accounts	Debit Card Information	25
	Deposit Rate Comparison	35
Loans	Auto Loan Prequalification	25
	Vehicle Search	35
Business	Business Credit Card Exploration	25
Digital Service	Digital Banking Features	25
General	Fraud Protection	25

As shown in Table 3, all three variants improve recall over the baseline, confirming that pre-login session representations carry signal beyond coarse page-visit counts. The bidirectional + mean pooling variant achieves the best overall performance, with a 13.38% Log Loss reduction and 1.88% macro Recall@1 gain. The advantage of bidirectional attention is intuitive: because the complete pre-login session is available at inference time, each event can attend to both preceding and subsequent actions, producing richer contextual representations than the causal setting where each event only sees its past. Similarly, mean pooling outperforms last-hidden-state pooling because the final event in a session can be a low-signal action such as a redirect or bounce; averaging over all event representations avoids the potential last-token bias.

4.3 Intent Taxonomy and Conversion Prediction Results

Table 4 shows example intents generated by the taxonomy pipeline at two granularity levels: I=25 and I=35. The iterative process produces semantically coherent categories, each grounded in observed clickstream patterns rather than predefined rules. After human review, we select I=35 for the following experiments. For clustering-based sampling, we select K-Means with K=50 based on silhouette score (0.535), Davies-Bouldin Index (1.478), and Calinski-Harabasz Index (19,250), indicating well-separated clusters suitable for stratified sampling.

To validate LLM label quality, we conduct a manual review on a stratified random sample spanning all intent categories. A human

Table 5: Conversion prediction results (micro metrics). Percentage changes are relative to the LLM teacher baseline (C).

	Model	Dim	Prec.	Rec.	F1
Hand-crafted features					
A	Site Retargeting Features	9	−41.3%	−23.6%	−35.8%
Intent-based representations					
B	Distilled (student)	35	−8.9%	−3.6%	−7.0%
C	LLM intent (teacher)	35	—	—	—
Self-supervised embedding					
D	**Session emb.**	**64**	**+7.3%**	−0.5%	**+4.3%**

annotator independently assesses whether the LLM-assigned labels correctly capture the session's browsing intent, rating each as *agree*, *partially agree*, or *disagree*. The strict agreement rate (full agree) is 97.0% and the lenient rate (agree + partially agree) is 99.5%, with the single disagreement involving a low-confidence edge case where no suitable taxonomy label existed. These results confirm that LLM-based label assignment is reliable enough to serve as a teacher signal for distillation without requiring exhaustive human annotation.

We report results using micro metrics for the following experiments. Unlike tile ranking, where rare clicks can drive high revenue and thus necessitate macro metrics, conversion frequency directly correlates with business impact, making frequency-weighted micro metrics the ideal evaluation for high-volume actions like credit card applications. The distilled student MLP achieves micro F1 of 0.822 against LLM-generated labels on held-out test data. This confirms that the lightweight classifier reproduces the teacher's intent assignments while operating directly on the 64-dimensional embedding.

To validate the utility of our representations, we evaluate four feature sets on the user conversion prediction task described in Section 4.1. All models use the same MLP architecture with binary cross-entropy loss, per-class positive weighting, and early stopping on validation loss, so differences in performance are attributable solely to input features:

- **A. Site retargeting features (SRF):** 9 hand-crafted page-category count features aggregated over rolling time windows.
- **B. Distilled intents (student):** 35-dimensional probability vector from the distilled intent classifier (Section 3.3).
- **C. LLM intents (teacher):** 35-dimensional raw intent confidence vectors produced by the LLM during label assignment (Section 3.3).
- **D. Session embedding:** 64-dimensional embedding (Section 3.2).

Hand-crafted features are insufficient. SRF (A) shows a 35.8% degradation in micro F1 relative to the LLM teacher, confirming that coarse page-category counts over rolling time windows fail to capture the behavioral nuances needed for conversion prediction. This motivates learning richer representations from the raw clickstream.

LLM teacher validates the taxonomy quality. The LLM teacher (C) serves as the reference baseline for intent-based representations. Its 35-dimensional confidence vector, despite being highly sparse, encodes meaningful predictive information derived from the full clickstream context.

Distillation effectively transfers teacher knowledge. The distilled student (B) retains 93% of the LLM teacher's predictive power, with only a 7.0% relative decrease in micro F1. This validates that knowledge distillation successfully transfers the majority of the teacher's signal into a lightweight MLP operating directly on the 64-dimensional embedding. The remaining gap reflects the inherent compression from per-session LLM judgments to a fixed classifier learned from a limited labeled sample, a necessary tradeoff for scaling from thousands of LLM calls to millions of MLP inferences (Section 4.4). Notably, combining distilled intents with SRF (A+B) yields +37% relative F1 improvement over SRF alone, confirming that intent labels capture substantially richer signal than count-based features.

Session embedding contains the richest signal. The session embedding (D) achieves the best micro F1, a 4.3% improvement over the LLM teacher, despite being only 64-dimensional. This is expected: the session embedding is the *input* to the distillation pipeline, and it encodes the full multi-modal clickstream in a dense representation. The distillation pipeline then extracts interpretable intent labels from this same embedding space, necessarily losing some information in exchange for human readability.

Together, these results validate the dual-output design: the session embedding serves quantitative downstream tasks with maximum predictive power, while the distilled intent labels, derived from the same embedding space, add interpretability, segmentation capability, and stakeholder-facing explanations at minimal cost to accuracy.

4.4 Inference Efficiency

A key motivation for knowledge distillation is production scalability. Assigning intent labels to raw clickstreams via direct LLM inference, either through an API or a self-hosted model, requires generating a full response per session, each incurring 2–4 seconds of latency. For example, at a scale of 100K sessions, this amounts to over 80 hours of serial inference time, which is infeasible for deployment.

By contrast, our pipeline from raw clickstreams to intent labels is extremely fast: the Transformer encodes 100K sessions into embeddings in 60 seconds, and the distilled MLP produces intent labels from those embeddings in 0.009 seconds—totaling under 61 seconds per 100K sessions on average. This represents a throughput improvement of more than 3,000× with only a 7% drop in labeling performance, reducing the high cost of large-scale data batch processing and retaining a high-quality production feature without separate text encodings or LLM inference.

5 Conclusion and Future Work

We presented a scalable system that produces both session intent embeddings and interpretable intent labels from web clickstream data, bridging the gap between raw behavioral signals and semantic understanding of user intent in financial services. The session embeddings outperform baselines on the home screen tile ranking task, demonstrating that they capture richer signals than aggregation-based features. The intent taxonomy provides actionable categories

that can be used for user segmentation, personalization, and marketing strategy. The distilled classifier also outperforms the handcrafted features while producing human-readable intent labels and enabling scalable inference at serving time, though it achieves slightly lower performance than the raw session embeddings, which is an expected tradeoff.

Limitations. We acknowledge that the system is built and evaluated on financial services clickstream data from a single institution; generalizability to other domains and downstream tasks remains to be validated.

Future work. First, we plan to conduct online A/B testing to measure the impact of session embeddings and intent labels on the home tile ranking task. Second, we aim to automate taxonomy refinement over time, enabling the intent catalog to adapt as user behavior and product offerings evolve. Finally, we plan to extend from session-level to user-level intent prediction by aggregating embeddings and intent labels across multiple sessions per user over a longer time period.

References

[1] Shelby D. Bernhard, Carson K. Leung, Vanessa J. Reimer, and Joshua Westlake. 2016. Clickstream Prediction Using Sequential Stream Mining Techniques with Markov Chains. In *Proceedings of the 20th International Database Engineering & Applications Symposium* (Montreal, QC, Canada) *(IDEAS '16)*. Association for Computing Machinery, New York, NY, USA, 24–33. doi:10.1145/2938503.2938535

[2] William Black, Alexander Manlove, Jack Pennington, Andrea Marchini, Ercument Ilhan, and Vilda Markeviciute. 2024. TRACE: Transformer-based user Representations from Attributed Clickstream Event sequences.

[3] David M Blei, Thomas L Griffiths, and Michael I Jordan. 2010. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)* 57, 2 (2010), 1–30.

[4] Andrei Broder. 2002. A taxonomy of web search. 36, 2 (2002), 3–10.

[5] Qiwei Chen, Huan Zhao, Wei Li, Pipei Huang, and Wenwu Ou. 2019. Behavior sequence transformer for e-commerce recommendation in alibaba. In *Proceedings of the 1st international workshop on deep learning practice for high-dimensional sparse data*. 1–4.

[6] Yongjun Chen, Zhiwei Liu, Jia Li, Julian McAuley, and Caiming Xiong. 2022. Intent contrastive learning for sequential recommendation. In *Proceedings of the ACM web conference 2022*. 2172–2182.

[7] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 120, 30 (2023), e2305016120.

[8] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based recommendations with recurrent neural networks. In *Proceedings of the 4th International Conference on Learning Representations (ICLR)*.

[9] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*. 8003–8017.

[10] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*. IEEE, 197–206.

[11] Dwipam Katariya, Juan Manuel Origgi, Yage Wang, and Thomas Caputo. 2024. Timesync: Temporal intent modelling with synchronized context encodings for financial service applications. *arXiv preprint arXiv:2410.12825*.

[12] Yong Soo Kim and Bong-Jin Yum. 2011. Recommender system based on click stream data using association rule mining. *Expert Syst. Appl.* 38, 10 (Sept. 2011), 13320–13327. doi:10.1016/j.eswa.2011.04.154

[13] Chao Li, Zhiyuan Liu, Mengmeng Wu, Yuchi Xu, Huan Zhao, Pipei Huang, Guoliang Kang, Qiwei Chen, Wei Li, and Dik Lun Lee. 2019. Multi-interest network with dynamic routing for recommendation at Tmall. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 2615–2623.

[14] Yueqing Liang, Liangwei Yang, Chen Wang, Xiongxiao Xu, Philip S Yu, and Kai Shu. 2025. Taxonomy-guided zero-shot recommendations with llms. (2025), 1520–1530.

[15] Erdi Olmezogullari and Mehmet S. Aktas. 2020. Representation of Click-Stream DataSequences for Learning User Navigational Behavior by Using Embeddings. In *2020 IEEE International Conference on Big Data (Big Data)*. 3173–3179. doi:10.1109/BigData50022.2020.9378437

[16] Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. TopicGPT: A prompt-based topic modeling framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2956–2984.

[17] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. 3982–3992.

[18] Chirag Shah, Ryen White, Reid Andersen, Georg Buscher, Scott Counts, Sarkar Das, Ali Montazer, Sathish Manivannan, Jennifer Neville, Nagu Rangan, et al. 2025. Using large language models to generate, validate, and apply user intent taxonomies. *ACM Transactions on the Web* 19, 3, 1–29.

[19] Alex Stein, Samuel Sharpe, Doron Bergman, Senthil Kumar, C Bayan Bruss, John Dickerson, Tom Goldstein, and Micah Goldblum. 2024. A simple baseline for predicting events with auto-regressive tabular transformers. *arXiv preprint arXiv:2410.10648* (2024).

[20] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1441–1450.

[21] Mengting Wan, Tara Safavi, Sujay Kumar Jauhar, Yujin Kim, Scott Counts, Jennifer Neville, Siddharth Suri, Chirag Shah, Ryen W White, Longqi Yang, et al. 2024. Tnt-llm: Text mining at scale with large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5836–5847.

[22] Gang Wang, Xinyi Zhang, Shiliang Tang, Haitao Zheng, and Ben Y Zhao. 2016. Unsupervised clickstream clustering for user behavior analysis. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 225–236.

[23] Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 4195–4205.

[24] Zihan Wang, Jingbo Shang, and Ruiqi Zhong. 2023. Goal-driven explainable clustering via language descriptions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 10626–10649.

[25] Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. 2025. Organize the web: Constructing domains enhances pre-training data curation. *arXiv preprint arXiv:2502.10341*.

[26] Xue Xia, Pong Eksombatchai, Nikil Pancha, Dhruvil Deven Badani, Po-Wei Wang, Neng Gu, Saurabh Vishwas Joshi, Nazanin Farahpour, Zhiyuan Zhang, and Andrew Zhai. 2023. Transact: Transformer-based realtime user action model for recommendation at pinterest. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5249–5259.

[27] Chao Zhang, Fangbo Tao, Xiusi Chen, Jiaming Shen, Meng Jiang, Brian Sadler, Michelle Vanni, and Jiawei Han. 2018. Taxogen: Unsupervised topic taxonomy construction by adaptive term embedding and clustering. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 2701–2709.

[28] Yuwei Zhang, Zihan Wang, and Jingbo Shang. 2023. Clusterllm: Large language models as a guide for text clustering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 13903–13920.

[29] Han Zhu, Xiang Li, Pengye Zhang, Guozheng Li, Jie He, Han Li, and Kun Gai. 2018. Learning tree-based deep model for recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1079–1088.