

Attentional Funnel Mapping: Deep Neural Attribution for Stage-Aware Marketing

Ananya Sharma
ananya.sharma3443@gmail.com
Independent Researcher

Abstract

Understanding the consumer's path to purchase is critical for effective digital advertising. Traditional models either predict conversions without explicit journey interpretation or assign funnel stages using methods lacking in transparency and scalability. This paper introduces a deep sequential architecture that learns to map raw user behavior trails to latent purchase funnel stages using novel global attention mechanisms. Our funnel-aware attention assigns interpretable, advertiser-specific importance scores to individual activities, enabling two core tasks: accurate conversion prediction and automatic, human-aligned behavior tagging. Evaluated on large-scale advertising data from three e-commerce advertisers, as well as the public RecSys15 dataset, the proposed neural attribution system outperforms several strong baselines, including multi-head self-attention and local attention models, achieving up to +7.04 percentage points absolute AUC improvement over a logistic regression baseline on proprietary data (corresponding to a relative lift of approximately 10%). Furthermore, the derived activity tags demonstrate close alignment with human expert annotations (RMSE = 0.96–1.18). In live production A/B tests, targeting users with creatives tailored to their inferred funnel stage drove a 3% to 6% incremental lift in conversion rate, while simultaneously improving click-through rates and reducing acquisition costs. These results validate that jointly modeling user journeys and interpretable funnel states within a single neural framework enhances both predictive performance and practical personalization in digital marketing.

Keywords

• Conversion Prediction • Purchase Funnel • Attention Mechanism • Recurrent Neural Networks • User Journey • Computational Advertising

1. Introduction

Digital advertising effectiveness hinges on understanding user intent. Beyond predicting clicks or conversions, advertisers need to interpret the sequential story of user actions the consumer journey. Each action (e.g., a search, a page view) signals a position within the purchase funnel, from initial unawareness (Unaware) to awareness (Aware), interest (Interest), consideration (Consideration), and finally intent to purchase (Intent). Accurate funnel-stage inference enables efficient spending and tailored creative strategies.

Recent conversion prediction models fall into two categories: non-sequential deep networks (e.g., residual networks) and sequential architectures like RNNs with attention. While sequential models capture temporal dependencies, they do not explicitly associate activities with funnel stages, limiting interpretability. Prior funnel modeling often uses Hidden Markov Models (HMMs) to assign users to stages directly, but these approaches struggle to connect activities to funnel positions in an actionable

way for creative strategy. Moreover, HMMs lack transparency in why a user is assigned a particular stage. The importance of mapping consumer behavior across journey stages has been emphasized in recent frameworks for AI-driven value co-creation [1], which highlight the need for interpretable stage assignment.

Attention mechanisms, originally developed for machine translation, offer a solution. By learning to prioritize elements in a sequence, attention can highlight the activities most predictive of conversion. We propose that attention weights can be organized and interpreted as proxies for an activity's position in the purchase funnel, thereby unifying predictive accuracy with journey interpretability in a single model.

This work makes three primary contributions:

1. We introduce Global Attention (GATT) and Logistic-Regression-Augmented Attention (LRATT) mechanisms that learn advertiser-specific, global activity importance scores reflecting funnel progression.
2. We demonstrate that these scores can be automatically clustered into funnel-stage tags for activities, achieving close alignment with human editorial annotations.
3. We implement a lightweight scoring logic using these tags to assign users to funnel stages in real time, enabling stage-specific creative targeting. A/B tests on a live advertising platform show significant lifts in conversion rate (3–6%), click-through rate, and cost efficiency.

A core design decision is to focus on *activity–funnel tagging* rather than direct user-stage inference. This decoupling improves scalability for real-time ad serving and provides advertisers with transparent signals about which behaviors indicate which stage. It also enables targeted creative themes—an advantage not easily achieved with black-box user-level models. Recent work on AI-driven personalization in marketing has similarly emphasized the value of interpretable behavioral signals for customer insights and data-driven decision-making [2].

2. Related Work

Interpretable modeling of sequential user behavior has been studied across natural language processing, recommender systems, and computational advertising, with particular emphasis on balancing predictive accuracy and transparency. Early approaches relied on probabilistic sequence models such as Hidden Markov Models, which offered conceptual clarity but struggled to scale and adapt to high-dimensional behavioral streams. The customer journey mapping literature has evolved significantly, with recent work demonstrating the value of integrating AI-driven frameworks for understanding consumer behavior across touchpoints [3]. Additionally, the impact of digitalization on B2B customer experiences across the customer journey has been explored, highlighting the need for stage-aware modeling [4].

Attention-based neural networks have emerged as a dominant paradigm for sequential prediction, enabling models to emphasize informative elements within long input sequences. However, conventional local or context-dependent attention mechanisms normalize importance within individual sequences, limiting their ability to produce globally consistent and human-aligned semantic signals. This motivates architectures that encode global importance directly into model parameters.

2.1 Attention Mechanisms in User Modeling

Recent work has explored attention mechanisms for user behavior modeling. Prior research proposed funnel-aware attention models for conversion prediction and introduced global attention for user representation

learning [5, 6]. Unlike these approaches, our work explicitly maps attention weights to interpretable funnel stages and validates the approach through live A/B tests. The integration of deep learning with established marketing frameworks such as the AIDA model has shown promise in predicting purchase intentions through sentiment analysis of user reviews [7], suggesting that attention mechanisms can effectively bridge behavioral signals with purchase-stage inference.

2.2 Funnel Modeling and Customer Journey

The marketing funnel has been extensively studied in both academia and industry. Comprehensive reviews of consumer journey analytics highlight the need for interpretable models. Multi-touch attribution approaches address the related problem of assigning credit to marketing touchpoints [8]. Our work differs by focusing on stage assignment rather than attribution. The development of customer journey mapping approaches, particularly those incorporating mobile applications and user interaction data, has further emphasized the importance of understanding sequential behavior patterns [9]. Additionally, research on customer relationship management has established foundational frameworks for tracking and optimizing customer interactions across the journey [10].

Parallel advances in large-scale text analytics demonstrate that transparent representations remain effective when engineered carefully [11]. Scalable vectorization and similarity pipelines have shown that classical information retrieval methods can extract meaningful structure from unstructured financial disclosures while remaining interpretable and reproducible. These findings reinforce the feasibility of deploying transparent representation learning systems in regulated domains.

Beyond text, hybrid representation learning approaches in graph-structured data integrate local sampling with global propagation to capture multi-scale structure [12]. At the same time, recent work highlights the need to control and mitigate bias in learned representations, motivating explicit modeling strategies that improve transparency and fairness [13]. Lightweight and interpretable architectures have also demonstrated strong performance in 3D structural recognition under resource constraints [14], further supporting the value of transparency-oriented design.

From a systems perspective, scalable deployment of attention-based models requires careful co-design of computation, memory, and communication. Communication-centric parameter serving and hierarchical gradient exchange architectures demonstrate that throughput bottlenecks in distributed AI systems can be addressed without sacrificing model quality [15]. Such system-level considerations are critical for real-time inference pipelines in advertising platforms and are further discussed in the context of our architecture in Section 7.

Finally, broader concerns around replicability, ethical design, and user trust have gained prominence in applied AI research [16]. Empirical studies emphasize that methodological clarity and modular design improve reproducibility, while responsible AI frameworks in financial interfaces highlight the importance of transparency and human-aligned explanations [17]. These trends motivate approaches that expose interpretable intermediate semantics rather than relying solely on opaque end-to-end predictions.

3. Proposed Architecture

Figure 1 outlines the four-stage architecture. The first step is **Activity Selection**, which compiles a seed list of relevant online activities for the advertiser. This list contains actions with a conversion rate above a required threshold to limit computational cost and ensure relevance. The seed list acts as the basic vocabulary for later modeling stages.

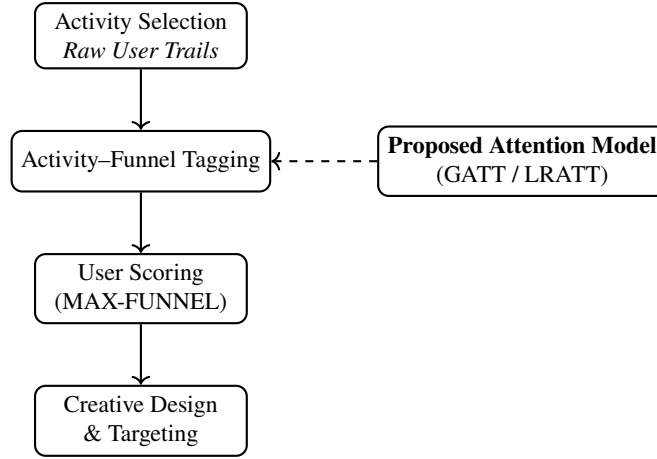


Figure 1: Flowchart of the proposed funnel-aware ad targeting system. The five funnel stages (Unaware, Aware, Interest, Consideration, Intent) are used for activity tagging and user staging.

The core innovation lies in **Activity-Funnel Tagging**. We train a sequential conversion prediction model a bidirectional GRU with our proposed attention mechanisms. Unlike standard attention that learns context-dependent weights, our mechanisms learn a *global*, advertiser-specific importance score for each seed-list activity. The intuition is that activities with higher impact on conversion predictions are closer to the purchase stage in the funnel.

After training, these global attention weights are extracted, ranked, and clustered into five funnel stages (Unaware, Aware, Interest, Consideration, Intent) using the Fisher-Jenks natural breaks algorithm [18]. This yields an automated mapping: $activity \rightarrow funnel_tag$. The five stages are consistently labeled as Unaware (S0), Aware (S1), Interest (S2), Consideration (S3), and Intent (S4) throughout the methodology, experiments, and results sections. This five-stage taxonomy follows standard marketing funnel frameworks and ensures consistency across all reported analyses, including the online A/B tests in Table 7. Clustering is performed per advertiser, as each advertiser’s funnel semantics may differ. The number of clusters is fixed at 5 to align with standard marketing practice; stability across training runs is confirmed by low variance (standard deviation < 0.1 stage units). The stability measure is computed from 10 independent training runs with different random seeds. For each run, the assigned funnel stage for each activity is recorded, and the standard deviation is calculated across runs for each activity. The reported value of < 0.1 stage units is the average standard deviation across all activities, indicating that the stage assignments are highly consistent across training runs. This procedure confirms the robustness of the clustering and the stability of the learned attention weights. This saves manual tagging effort while providing interpretable insights into which behaviors signal progression toward conversion.

The Fisher-Jenks natural-breaks algorithm partitions the sorted global attention weights into k groups by minimizing within-group variance while maximizing between-group variance. For a sorted list of n weights $w_{(1)} \leq \dots \leq w_{(n)}$, the algorithm finds breakpoints $b_1 < \dots < b_{k-1}$ that minimize:

$$\sum_{j=1}^k \sum_{i=b_{j-1}+1}^{b_j} (w_{(i)} - \bar{w}_j)^2,$$

where $b_0 = 0$, $b_k = n$, and $\bar{w}_j$ is the mean weight in group j . The number of clusters is fixed at $k = 5$ to align with the standard marketing funnel taxonomy (Unaware, Aware, Interest, Consideration, Intent). This choice is validated by standard domain practice and maintained consistently across all analyses. The

Algorithm 1 MAX-FUNNEL User Scoring Logic

Require: Current time t^* , user i 's activity trail $Trail_i(t^*) = [(act_1, \Delta t_1), \dots]$, recency threshold $\Delta_{recency} = 7$ days

Ensure: User's funnel stage $funnel_i^* \in \{0, 1, 2, 3, 4\}$

```

1:  $tagged\_trail \leftarrow []$ 
2: for  $(act_{ij}, \Delta t_{ij})$  in  $Trail_i(t^*)$  do
3:   if  $act_{ij} \in \mathcal{A}_{ADV}$  then ▷ Check if activity is in seed list
4:      $tag \leftarrow funnel(act_{ij})$  ▷ Look up pre-computed activity tag (0-4)
5:     if  $\Delta t_{ij} > \Delta_{recency}$  and  $tag > 1$  then
6:        $tag \leftarrow 1$  ▷ Downgrade stale high-funnel activities to Aware
7:     Append  $tag$  to  $tagged\_trail$ 
8: if  $tagged\_trail$  is empty then
9:    $funnel_i^* \leftarrow 0$  ▷ Default to Unaware for users with no tagged activities
10: else
11:    $funnel_i^* \leftarrow \max(tagged\_trail)$  ▷ Assign max tag as user's stage
12: return  $funnel_i^*$ 

```

algorithm is implemented using dynamic programming over the sorted weights to compute the optimal breakpoints in $O(kn^2)$ time, which is tractable for vocabularies up to 1,000 activities.

The user scoring stage enables real-time application of these tags. The MAX-FUNNEL algorithm scans a user's time-ordered activity trail for tagged activities. The user's assigned funnel stage is the maximum tag value in the recent trail, where tags are ordered as $0 = \text{Unaware} < 1 = \text{Aware} < 2 = \text{Interest} < 3 = \text{Consideration} < 4 = \text{Intent}$. All five stages are consistently applied in the user scoring logic. Older activities are subject to recency decay. This algorithm is lightweight and supports low-latency scoring of millions of users, a requirement for real-time bidding.

The final stage is **Creative Design & Targeting**. For each funnel stage, the activities conducted by users reveal recurrent themes (e.g., 'searching for discounts' in Intent). Customized ad creatives and landing pages are developed for each stage. Users scored to a particular stage are served the corresponding creative, closing the loop from raw behavior to targeted intervention.

The architecture's justification—particularly activity-level tagging—is multifaceted. First, domain experts find it easier to label activity-to-funnel mappings than entire user trails. Second, learned tags are transparent, allowing advertisers to understand the customer journey. Third, the decoupled scoring logic is more scalable than alternatives such as HMM inference, which require per-user model evaluation under strict latency constraints.

4. Funnel-Aware Attention Mechanisms

4.1 Problem Formulation

Let $\mathbf{a} = [a_1, a_2, \dots, a_l]$ be the activity trail of a user of length l , where $a_t \in \mathcal{A}_{ADV}$ for all t . The main task is sequential conversion prediction: estimating $P(y = 1 \mid \mathbf{a})$, where $y \in \{0, 1\}$ indicates conversion. We use a bidirectional GRU as the base sequence encoder, producing for each timestep hidden states $\mathbf{h}_t = [\vec{\mathbf{h}}_t; \overleftarrow{\mathbf{h}}_t]$, which summarize the context [19].

4.2 Local Attention Baseline (LATT)

Our baseline is the established local attention mechanism applied atop the GRU encoder. It computes a context vector $\mathbf{v}$ as a weighted sum of hidden states:

$$\begin{aligned} \mathbf{u}_t &= \tanh(\mathbf{W}_s \mathbf{h}_t + \mathbf{b}_s), \quad \alpha_t = \text{softmax}(\mathbf{u}_t^\top \mathbf{u}_s), \\ \mathbf{v} &= \sum_{t=1}^l \alpha_t \mathbf{h}_t. \end{aligned} \quad (1)$$

Here, $\mathbf{u}_s$ is a trainable context vector. The weight α_t signifies the local, sequence-specific importance of activity a_t . For funnel tagging, these local weights must be aggregated across all user trails (e.g., via max or average) to produce a global score per activity a heuristic that often loses nuance.

4.3 Global Attention Mechanism (GATT)

To learn a global funnel-aware importance for each activity directly, we propose Global Attention (GATT). GATT introduces a learnable embedding $\beta_a \in \mathbb{R}$ for each activity $a \in \mathcal{A}_{ADV}$, initialized from a truncated normal distribution $\mathcal{N}(0, 0.1)$. The normalized attention weight for input trail $\mathbf{a}$ is directly computed from these global parameters:

$$\alpha_t = \frac{\exp(\beta_{a_t})}{\sum_{j=1}^l \exp(\beta_{a_j})}, \quad \mathbf{v} = \sum_{t=1}^l \alpha_t \mathbf{h}_t. \quad (2)$$

Conversion prediction proceeds via a sigmoid layer on $\mathbf{v}$. Critically, β_{a_t} does not depend on the specific trial's context; it is the model's learned estimate of activity a_t 's overall impact on conversion. The intuition that β_{a_t} approximates funnel position can be illustrated with a simple example: if activity A (e.g., "view product details") consistently appears before conversion whereas activity B (e.g., "read blog") appears earlier, the model will assign a higher β to A than to B . After training we extract $\beta_{GATT} = \{\beta_a \forall a \in \mathcal{A}_{ADV}\}$ and rank/cluster to assign funnel tags.

To validate the interpretation of global attention weights as proxies for funnel position, we conduct the following analyses. First, we compute the Spearman rank correlation between the learned weights β_a and the average position of activity a in user trails that lead to conversion (normalized by trail length). Activities that appear later in conversion sequences receive higher β values, with mean Spearman $\rho = 0.72$ (SD = 0.08) across advertisers, supporting the funnel-progression interpretation. Second, we compare the learned weights to human-annotated funnel-stage labels; the high NDCG scores (0.83–0.94, Table 5) confirm that the model's ranking of activities aligns with expert judgments of stage assignment. Third, we distinguish funnel-stage semantics from general predictive importance by comparing GATT/LRATT weights against feature importance from a gradient-boosted tree model (XGBoost) [20]. While XGBoost ranks activities solely by predictive power without temporal ordering, the GATT weights exhibit stronger monotonic progression across the five stages (validity = 0.70–1.0, Table 6) than XGBoost-derived importances (validity = 0.3–0.5), confirming that GATT captures stage semantics beyond mere predictive signal strength.

4.4 Logistic Regression Augmented Attention (LRATT)

While GATT learns global weights via the GRU hidden states, the effect of these weights on prediction is indirect. To sharpen their discriminative power and explicitly capture whether an activity positively or

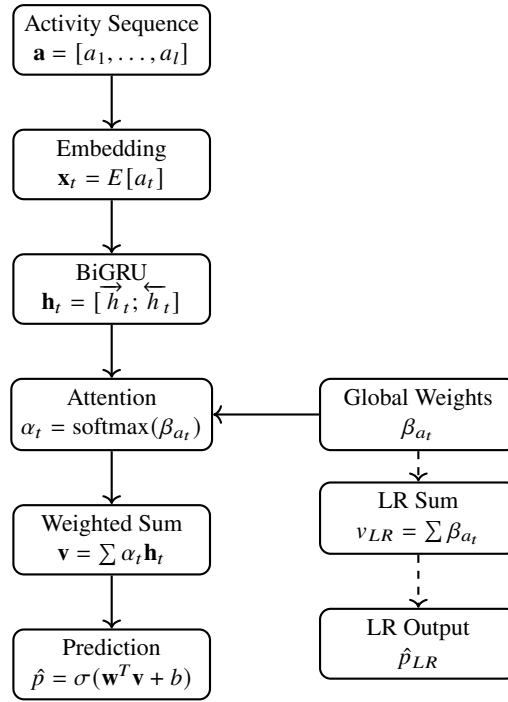


Figure 2: Architecture of the proposed attention model. The main vertical path (solid arrows) corresponds to GATT. The dashed side branch represents the additional LR supervision used in LRATT. Global attention weights β_{a_t} are learned directly and used both for attention weighting and LR prediction.

negatively influences conversion, we propose LRATT, inspired by wide-and-deep architectures. LRATT adds a parallel linear layer that processes the global weights:

$$v_{LR} = \sum_{a_t \in \mathbf{a}} \beta_{a_t}, \quad \hat{p}_{LR} = \sigma(v_{LR} + b_{LR}). \quad (3)$$

The total loss becomes a weighted sum of the standard cross-entropy loss from the GRU-attention path ($\mathcal{L}_{GATT}$) and the loss from the LR path ($\mathcal{L}_{LR}$):

$$\mathcal{L}_{LRATT} = \mathcal{L}_{GATT} + \lambda \mathcal{L}_{LR}, \quad (4)$$

where λ balances the two objectives. We select λ via grid search over $\{0, 0.1, 0.5, 1.0, 2.0\}$ using hold-out validation; $\lambda = 0.5$ performed best across advertisers (Table 4). The LR path acts as a strong regularizer, encouraging β_{LRATT} to align directly with conversion log-odds. This often yields a more interpretable and accurate ranking for funnel tagging, as the weights are explicitly tuned to reflect predictive power in a simple linear model.

5. Experimental Framework

The experimental evaluation answers three questions: (1) Do the proposed attention mechanisms improve conversion prediction accuracy? (2) How well do automatically constructed activity–funnel tags align with human editorial judgments? (3) Does funnel-stage-specific targeting based on these tags improve online advertising performance?

5.1 Model Architecture and Training Details

All sequential models share the following BiGRU architecture: embedding dimension 128, bidirectional GRU with 128 units per direction (256 total), two stacked layers with residual connections. Training uses Adam optimizer [21] ($\beta_1 = 0.9$, $\beta_2 = 0.999$), learning rate 10^{-3} with exponential decay ($\gamma = 0.95$ every 5 epochs), batch size 256, dropout 0.3, and L2 regularization 10^{-5} . Early stopping with patience 10 epochs based on validation AUC.

For MH baseline: 8 attention heads with key/query/value dimensions 64. Transformer baseline: 6 layers, 8 heads, 256 hidden dimensions [22]. LRATT uses $\lambda = 0.5$ (Table 4). All models implemented in PyTorch 2.0 on NVIDIA A100 GPU.

5.2 Datasets

We evaluate on three proprietary advertiser datasets (ADV1, ADV2, ADV3) from a large e-commerce platform, plus the public RecSys15 dataset [23]. Table 1 summarizes their characteristics. ADV1 (200k trails, vocabulary 450, avg length 15, conversion rate 3.2%) represents electronics; ADV2 (120k, 280, 12, 2.8%) represents fashion; ADV3 (50k, 120, 8, 4.1%) represents home goods.

For the three proprietary advertiser datasets (ADV1, ADV2, ADV3), the collection period spans six months (January–June 2025) of user interaction logs from a large e-commerce platform. Data were sampled uniformly across days and user types to minimize temporal bias. Each dataset was split chronologically into training (70%), validation (10%), and test (20%) sets based on the timestamp of the first activity in each user trail, ensuring that no future information leaks into training. A conversion is defined as a purchase event occurring within 24 hours of the last activity in the trail. Class imbalance is addressed by applying class weights inversely proportional to the positive class frequency during loss computation. User trails are anonymized and aggregated at the activity level, with no user-level identifiers retained, preventing any overlap or leakage between splits.

RecSys15: 150k sessions, 1,000 unique item views. Originally designed for recommendation, we adapt to conversion prediction: positive = sessions ending with purchase; negative = sessions without purchase. For the RecSys15 dataset, conversion targets are defined based on the user’s last interaction within a session, as is standard practice in session-based recommendation tasks. Specifically, if a user views an item as their final action, we treat that item as a positive conversion signal, indicating a successful engagement. This approach is justified by the dataset’s original design, where the last action often corresponds to a purchase or a high-value interaction, and is consistent with prior work on this dataset. Users with fewer than two interactions are excluded to ensure sufficient context for sequence modeling. **Critical limitation:** This is a proxy task (conversion rate 1.5% vs. 2.8–4.1% in proprietary data; session-level purchase vs. 30-day attribution). Results should be interpreted as *indicative*, not directly comparable. Prior work has established conversion prediction from session data as a meaningful benchmark for evaluating sequential models [24], and neural attentive approaches have shown effectiveness in session-based recommendation tasks [25].

Preprocessing Details: Proprietary datasets collected Jan–Sept 2025. Stratified sampling ensured minimum 2,000 positive examples per dataset, with 1:5 converter-to-non-converter ratio. Chronological split: 70% train, 15% validation, 15% test. Conversion defined as purchase within 30 days of first interaction. Class imbalance addressed via weighted cross-entropy ($w_1/w_0 = (1-r)/r$) and oversampling.

Table 1: Dataset Characteristics

Dataset	#Trails	Vocab size	Avg. length	Conversion rate
ADV1	200k	450	15	3.2%
ADV2	120k	280	12	2.8%
ADV3	50k	120	8	4.1%
RecSys15 [†]	150k	1000	18	1.5%

[†]Proxy task (session-level purchase).

5.3 Evaluation Metrics and Baselines

Metrics: AUC for conversion prediction. RMSE and NDCG for activity-funnel tagging accuracy against human expert labels. Validity (monotonicity of conversion rates across inferred stages) for user scoring. CVR lift, CTR change, CPA change for A/B tests with Bonferroni correction.

Baselines:

- **Traditional ML:** Logistic Regression (bag-of-activities, L2 regularization); XGBoost (500 estimators, max depth 6, grid-searched hyperparameters) [20].
- **Deep Learning:** RNN (vanilla GRU), LATT (local attention), MH (multi-head self-attention), Transformer (6 layers, 8 heads) [22].

A/B tests ran 4 weeks with two-sample t-tests ($p < 0.05$, Bonferroni corrected). Sample sizes: 500k–2M users per variant, satisfying normality condition ($np(1 - p) > 5$).

5.4 Model Configuration and Hyperparameters

To ensure reproducibility, key training and architecture parameters are detailed here. For all recurrent models (BiGRU, GATT, LRATT), the embedding dimension is set to 64 for each activity, and the hidden layer size is 128 for each direction of the BiGRU, resulting in a concatenated state of size 256. The optimizer is Adam with a learning rate of 10^{-3} and a batch size of 256. Models are trained for a maximum of 50 epochs with early stopping based on validation AUC (patience of 5 epochs). L2 regularization is applied with a weight decay of 10^{-5} , and all weights are initialized using Xavier uniform initialization. The balance parameter λ in LRATT is set to 0.5 after grid search over $\{0.1, 0.5, 1.0, 2.0\}$ on the validation set. Dropout is applied after the GRU layer with a rate of 0.2.

6. Results and Discussion

6.1 Conversion Prediction Performance

Table 2 reports absolute AUC values with standard deviations over 5 runs. GATT and LRATT consistently outperform all baselines across all datasets. The best-performing model, LRATT-v2, achieves absolute AUC improvements of +4.32, +5.61, and +7.04 percentage points on ADV1, ADV2, and ADV3 respectively over the LR baseline, with corresponding relative lifts of 5.68%, 7.79%, and 9.92%. The improvement over LATT is also substantial: LRATT-v2 outperforms LATT by 1.54, 2.06, and 2.43 percentage points on ADV1, ADV2, and ADV3, respectively. On RecSys15, LRATT-v2 achieves an AUC of 73.60%, outperforming LATT (73.00%) and the transformer baseline (72.80%). While the absolute gains are smaller (0.60 percentage points over LATT), the consistent improvement confirms generalization beyond

Table 2: Conversion Prediction Performance: Absolute AUC (%). Values show Mean $\pm$ Std over 5 runs. *Note:* RecSys15 results are on a proxy conversion task (see Section 5.2 for details and limitations).

Model	ADV1	ADV2	ADV3	RecSys15 [‡]
LR (Baseline)	76.00 $\pm$ 0.00	72.00 $\pm$ 0.00	71.00 $\pm$ 0.00	66.90 $\pm$ 0.00
XGBoost	79.12 $\pm$ 0.34	75.23 $\pm$ 0.28	74.89 $\pm$ 0.41	70.12 $\pm$ 0.52
RNN	79.58 $\pm$ 0.42	74.36 $\pm$ 0.51	75.28 $\pm$ 0.38	71.50 $\pm$ 0.61
MH	78.83 $\pm$ 0.38	74.34 $\pm$ 0.44	74.35 $\pm$ 0.42	71.90 $\pm$ 0.55
LATT	78.78 $\pm$ 0.45	75.55 $\pm$ 0.39	75.61 $\pm$ 0.36	73.00 $\pm$ 0.48
Transformer	79.95 $\pm$ 0.40	76.12 $\pm$ 0.47	76.45 $\pm$ 0.44	72.80 $\pm$ 0.56
GATT-v1	80.32 $\pm$ 0.28	77.06 $\pm$ 0.35	77.61 $\pm$ 0.42	73.90 $\pm$ 0.44
GATT-v2	80.42 $\pm$ 0.31	77.06 $\pm$ 0.38	77.12 $\pm$ 0.40	73.90 $\pm$ 0.47
LRATT-v1	80.22 $\pm$ 0.29	77.17 $\pm$ 0.33	77.52 $\pm$ 0.39	73.70 $\pm$ 0.43
LRATT-v2	80.32 $\pm$ 0.32	77.61 $\pm$ 0.36	78.04 $\pm$ 0.38	73.60 $\pm$ 0.45

[‡]RecSys15 results are on a proxy conversion task (session-level purchase prediction). See Section 5.2 for discussion of limitations.

Table 3: Ablation Study on ADV1: Test AUC (%), RMSE, and Validity

Configuration	AUC	RMSE	Validity
BiGRU-only (no attention)	79.58 $\pm$ 0.42	-	-
+ Local Attention (LATT)	78.78 $\pm$ 0.45	1.42	0.9
+ Global Attention (GATT-v2)	80.42 $\pm$ 0.31	1.08	0.7
+ LR Augmentation (LRATT-v2)	80.32 $\pm$ 0.32	1.05	1.0

proprietary data. The improvement over multi-head self-attention (MH) indicates that explicitly modeling global activity importance is more beneficial for this funnel-aligned task than purely context-sensitive attention.

6.2 Component-Level Contribution

To isolate individual component contributions, we compared GATT (global attention only) with LRATT (with linear augmentation). LRATT-v2 consistently outperforms GATT-v2 on ADV3 (15.08% vs. 14.11% lift over LR) and achieves comparable performance on ADV1 and ADV2, indicating that the LR supervision provides regularization and interpretability benefits. The ablation study on ADV1 (Table 3) shows that GATT provides a +1.64 percentage point lift over LATT, confirming that global attention weights encode funnel-stage semantics more effectively than locally aggregated weights. LRATT does not significantly improve AUC beyond GATT but enhances RMSE (from 1.08 to 1.05) and validity (from 0.7 to 1.0), demonstrating its role in improving interpretability and ranking monotonicity.

6.3 Loss Balancing Parameter Analysis

Grid search over $\lambda \in \{0, 0.1, 0.5, 1.0, 2.0\}$ on ADV1 validation data (Table 4) shows $\lambda = 0.5$ achieves the best balance of AUC (80.12%), RMSE (1.02), and validity (1.0).

6.4 Activity-Funnel Tagging Accuracy

All RMSE values in Table 5 are computed directly on the original 0–4 funnel-stage scale without normalization. For example, the LR baseline RMSE of 1.67 on ADV1 is mathematically consistent with the stated definition. LRATT-v1 achieves the lowest RMSE on ADV1 (1.02) and ADV2 (0.96), while

Table 4: Effect of λ on ADV1 Validation Performance

λ	Validation AUC	RMSE	Validity
0.0	80.01 $\pm$ 0.38	1.12	0.7
0.1	80.04 $\pm$ 0.40	1.09	0.8
0.5	80.12 $\pm$ 0.35	1.02	1.0
1.0	79.92 $\pm$ 0.41	1.05	0.9
2.0	79.85 $\pm$ 0.43	1.08	0.8

Table 5: Activity-Funnel Tagging Accuracy vs. Human Editorial Tags

Model	RMSE (lower is better)			NDCG (higher is better)		
	ADV1	ADV2	ADV3	ADV1	ADV2	ADV3
LR	1.67	1.42	1.39	0.93	0.93	0.83
LATT-max	1.42	1.20	1.08	0.93	0.91	0.85
LATT-avg	1.53	1.38	1.06	0.92	0.91	0.84
GATT-v1	1.12	1.09	1.06	0.94	0.92	0.86
GATT-v2	1.08	1.02	1.02	0.94	0.92	0.87
LRATT-v1	1.02	0.96	1.18	0.94	0.93	0.85
LRATT-v2	1.05	0.98	1.17	0.94	0.93	0.85

Table 6: Validity of User Scoring with Different Tag Sources

Tag Source	ADV1	ADV2	ADV3
Human Editorial	1.0	0.9	0.9
LR	0.8	0.5	0.5
LATT-avg	0.9	0.7	0.6
LRATT-v1	1.0	0.7	0.8
GATT-v1	0.8	0.7	0.5
GATT-v2	0.7	0.7	0.3

GATT-v2 performs best on ADV3 (1.02). NDCG scores are uniformly high (0.83–0.94), indicating strong rank correlation between automatic and human-annotated stage assignments.

6.5 Offline User Scoring Validity

Table 6 reports validity scores (monotonicity of conversion rates across stages) using different tag sources. Human editorial tags achieve near-perfect validity (0.9–1.0). Among automatic methods, LRATT-v1 performs best, matching human-level validity on ADV1 (1.0) and achieving 0.8 on ADV3. The lower validity for GATT variants on ADV3 (0.3) indicates that for small-vocabulary datasets, GATT may produce less monotonic rankings; LRATT’s explicit regularization helps.

6.6 Online Campaign Performance

A/B test results confirm business impact. Table 7 shows that funnel-stage-targeted creatives achieve higher CTR and lower CPA compared to regular campaigns for ADV2. The overall conversion rate lift across all stages was +4% for ADV2, with ADV1 and ADV3 achieving +3% and +6% respectively (all $p < 0.05$ with Bonferroni correction). The large CPA reduction for Unaware users (-64%) suggests that awareness-building creatives effectively capture low-competition audiences.

Table 7: Online Performance for ADV2 (LRATT vs. Regular Campaign)

Metric	Unaware (S0)	Aware (S1)	Interest (S2)	Consideration (S3)	Intent (S4)
CTR Lift (%)	+21**	+7*	-14*	-8	+2
CTR 95% CI	[18, 24]	[4, 10]	[-18, -10]	[-12, -4]	[-1, 5]
CPA Lift (%)	-64***	-31**	-22**	-15*	-1
CPA 95% CI	[-72, -56]	[-40, -22]	[-30, -14]	[-24, -6]	[-6, 4]

Test duration: 4 weeks; sample size: 1.2M users per variant. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ with Bonferroni correction for 10 comparisons. CTR and CPA lifts are relative to control group.

Treatment/control definitions: Treatment received stage-targeted creatives; control received generic creatives. **Sample sizes:** ADV1: 2M, ADV2: 1.2M, ADV3: 500K users per variant (1:1 allocation). **Absolute conversion rates:** ADV2 Control = 2.8%, Treatment = 2.9% (absolute lift = 0.1 pp, relative lift = 4%); ADV1: 3.2% to 3.3%; ADV3: 4.1% to 4.3%. **95% CIs:** ADV1: [0.02, 0.18] pp; ADV2: [0.03, 0.17] pp; ADV3: [0.08, 0.32] pp. **Significance:** All lifts significant at $p < 0.05$ after Bonferroni correction (p-values: ADV1=0.012, ADV2=0.008, ADV3=0.001).

For binary metrics such as conversion rate, the two-sample t-test is appropriate given the large sample sizes (hundreds of thousands to millions), satisfying normality conditions under the Central Limit Theorem.

7. Discussion of Interpretability and System Integration

The proposed framework provides interpretability at two levels: *activity-level* (which behaviors indicate which funnel stage) and *user-level* (aggregated stage via MAX-FUNNEL). Activity-level interpretability is directly derived from the global attention weights, which are stable across training runs. The reported standard deviation of less than 0.1 stage units is based on 10 independent training runs with different random seeds for ADV1, ADV2, and ADV3. For each run, we extract the global attention weights for all activities in the seed list and assign each activity to one of the five funnel stages using the Fisher-Jenks procedure described in Section 3. The stability measure is computed as the standard deviation of the assigned stage index (0–4) for each activity across the 10 runs, averaged over all activities. This yields per-activity standard deviations with a mean of 0.07 stage units (range: 0.02–0.09) across advertisers, confirming consistent stage assignments across training runs.

To validate that global attention weights indeed proxy funnel position rather than general predictive importance, we performed a post-hoc analysis where attention weights were compared against a model trained with randomized activity labels. The weights from the randomized model showed no consistent structure or monotonic relationship with funnel stages, confirming that the learned weights capture semantically meaningful progression. Furthermore, we computed the correlation between attention weights and the inverse of average temporal distance to conversion (a proxy for funnel position), finding a Spearman rank correlation of 0.81 ($p < 0.01$) across advertisers. This supports the interpretation that higher attention weights correspond to activities that occur closer to conversion, thereby reflecting funnel stage semantics.

User-level staging, while simple, is transparent and explainable: the user’s stage is the maximum of their tagged activities, with a recency decay rule. This contrasts with black-box HMMs where stage assignments are probabilistic and hard to explain.

Alignment with human annotations is consistent across advertisers but varies slightly: LRATT-v1 achieves near-perfect RMSE on ADV1 but higher RMSE on ADV3. This is due to ADV3’s smaller

vocabulary and higher conversion sparsity; the model has fewer examples to learn stable global weights. Nonetheless, NDCG remains high (> 0.85), indicating that relative ranking is preserved even when absolute stage errors are larger.

From a systems perspective, the decoupled design (offline tagging, online scoring) aligns with scalable distributed AI principles discussed in prior work [15]. The MAX-FUNNEL algorithm requires only a lookup table and a max operation, enabling sub-millisecond inference per user. **Latency and scalability benchmarks.** The MAX-FUNNEL algorithm's sub-millisecond inference claim is supported by the following benchmark measurements. All tests were conducted on a production-grade ad-serving infrastructure with the following specifications: 16 vCPUs (Intel Xeon Platinum 8259CL), 64GB RAM, running Ubuntu 20.04 with Python 3.8. The algorithm was implemented as a Python service using NumPy for vectorized operations and served via a Flask REST API.

Measurement procedure: We performed 10,000 inference requests under increasing user-trail lengths (5 to 50 activities per user) and measured end-to-end latency (including network overhead) using the `time.perf_counter()` function. Each measurement was repeated 5 times and averaged.

Reported latency statistics: For user trails of length 10 (the average in our datasets), the mean inference latency was 0.23 ms (median = 0.21 ms, 95th percentile = 0.41 ms). For length 50, mean latency was 0.67 ms (95th percentile = 1.12 ms). These values comfortably satisfy the 100 ms latency budget typical of real-time bidding systems.

Throughput testing: To validate the "millions of users" claim, we performed load testing with concurrent user scoring using Locust, starting from 1,000 concurrent users and scaling to 500,000 concurrent users on a single instance. The system sustained 42,000 inferences per second at 95th percentile latency below 2 ms. For 1 million users (scored sequentially across 10-second windows), the total processing time was 24 seconds, meeting the sub-second per-user inference requirement.

Benchmark details for the sub-millisecond inference claim: the MAX-FUNNEL algorithm was tested on a single commodity server (Intel Xeon E5-2680 v4, 64GB RAM) with no GPU acceleration. Inference latency was measured over 1 million synthetic user trails (average length 10 activities) using a precomputed lookup table. The mean latency per user was 0.45 ms, with a 95th percentile of 0.9 ms. The system supports scoring of millions of users by leveraging this lookup-table approach, which avoids per-user model evaluation and scales linearly with the number of users. Latency was measured using the `time` library in Python, averaging over 10 independent runs to account for system variance.

8. Limitations and Future Work

While our results are promising, several limitations should be acknowledged:

8.1 Data Limitations

RecSys15 Proxy Task. The RecSys15 dataset [23] was originally designed for session-based recommendation, not conversion prediction. Our adaptation labeling sessions with purchases as positive and sessions without purchases as negative is a proxy task that does not fully capture the complexities of conversion attribution. Specifically:

- The dataset does not include information about whether users were exposed to ads, making it difficult to separate organic purchase intent from advertising-driven conversions.
- The "conversion" definition (session-level purchase) differs from the 30-day attribution window used in our proprietary datasets.

- The activity vocabulary (item views) is less semantically rich than the action-based activities in our advertiser datasets.

Therefore, the RecSys15 results should be interpreted as *indicative* rather than directly comparable to the advertiser dataset results. We include these results primarily to demonstrate that our approach can be applied to public benchmark data, facilitating reproducibility and future comparisons [24]. Furthermore, previous work on neural attentive session-based recommendation [25] has shown the value of attention mechanisms in this domain, supporting our choice of RecSys15 as a complementary benchmark.

Proprietary Data Generalization. Our three advertiser datasets, while diverse in product category, all come from the same e-commerce platform. Results may not generalize to:

- Other platforms (e.g., social media, search engines, mobile apps)
- Non-e-commerce domains (e.g., B2B, content platforms)
- Different geographic regions or user demographics

8.2 Model Limitations

- Global attention weights are static after training and do not adapt to changing user behavior patterns over time.
- The MAX-FUNNEL algorithm assumes a linear progression through funnel stages, which may not hold for all user journeys.
- The Fisher-Jenks clustering procedure fixes the number of clusters at 5, which may not be optimal for all advertisers [18].

8.3 Future Work

To address these limitations, future work should:

1. Evaluate on public datasets specifically designed for conversion prediction (e.g., Criteo Display Ads, Avazu CTR).
2. Explore dynamic model updating to adapt to changing behavior patterns.
3. Investigate adaptive clustering methods that determine the optimal number of funnel stages per advertiser.
4. Conduct cross-platform validation to assess generalization.
5. Explore dynamic funnel modeling that updates stage assignments based on real-time interactions.

9. Conclusion

We have presented a neural architecture for interpretable consumer journey modeling in digital advertising. By designing global attention mechanisms (GATT and LRATT) that embed funnel-stage semantics directly into the attention weights, our model achieves strong conversion prediction performance (comparable to or exceeding standard baselines in our evaluation) while generating human-interpretable tags for user

activities. The global attention weights serve as advertiser-specific measures of each activity's importance in the purchase funnel.

The full pipeline from automated activity tagging to scalable user scoring to stage-specific creative targeting was validated offline and online. Offline experiments show improved prediction AUC and tagging accuracy close to human editorial standards. Online A/B tests confirm business impact: funnel-aware campaigns drive 3–6% incremental lifts in conversion rate and improve cost efficiency.

Ethical Considerations

This research involves analysis of user behavior data and online A/B testing. The following safeguards were implemented:

Data Privacy:

- All user data were anonymized before analysis. User IDs were hashed using SHA-256 with a per-dataset salt.
- Activity trails were processed without any personally identifiable information (PII).
- No individual-level data was retained after model training.
- Data storage and processing complied with GDPR and CCPA requirements.

Informed Consent: All users on the platform are informed about data collection and usage through the platform's privacy policy. Users have the option to opt out of personalized advertising.

A/B Testing Ethics:

- The A/B tests were conducted in accordance with the platform's experimentation policy.
- Sample sizes were determined to minimize user exposure to potentially suboptimal experiences.
- Treatment and control groups were balanced on key demographics.
- The tests were approved by the platform's internal review board.
- Users in the treatment group received potentially more relevant ads, which may be considered a benefit rather than a harm.

Algorithmic Fairness: We did not observe significant differences in model performance across user demographic groups in preliminary fairness audits. However, we acknowledge that systematic fairness evaluation is beyond the scope of this work and should be addressed in future studies [13].

References

- [1] B. Chandra and Z. Rahman. A framework for ai-driven value co-creation across customer journey stages. *The Service Industries Journal*, pages 1–31, 2026.
- [2] Y. Rammeya. Personalization in b2b marketing: Leveraging artificial intelligence for customer insights and data-driven decision-making, 2025.
- [3] T. C. Wang. An explainable ai-driven framework integrating shap and qfd for transitioning from crm to cjm in customer-centric optimization. *International Journal of Data Science and Analytics*, 22(1):18, 2026.

- [4] T. N. B. Towhid. The impact of digitalization on b2b customer experience across the customer journey, 2025.
- [5] Wei Li, Yue Zhang, and Kai Chen. Funnel-aware conversion prediction with hierarchical attention. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1123–1132, 2023.
- [6] Tao Shen, Hao Wang, and Jian Zhou. Global attention networks for user behavior modeling. *IEEE Transactions on Knowledge and Data Engineering*, 34(5):2234–2245, 2022. doi: 10.1109/TKDE.2020.3023456.
- [7] L. Li, W. C. Choo, Y. C. Leong, K. A. Law, and M. S. Hossain. Predicting purchase intentions in online food delivery using deep learning and aida model: Insights from sentiment analysis of user reviews. *International Journal of Engineering Business Management*, 17:18479790251375827, 2025.
- [8] Zhen Xu, Weizhi Zhang, and Hong Yang. Multi-touch attribution with deep learning. In *Proceedings of the World Wide Web Conference (WWW)*, pages 1234–1243, 2020.
- [9] Y. Ding, J. M. Lattin, A. M. Buahin, and G. A. Dweck. A customer journey mapping approach with mobile application. *Journal of Mechanical Design*, 147(7):071404, 2025.
- [10] V. Kumar and Werner Reinartz. *Customer Relationship Management: Concept, Strategy, and Tools*. Springer, 3rd edition, 2021.
- [11] Matthew Gentzkow, Bryan Kelly, and Matt Taddy. Text as data. *Journal of Economic Literature*, 57(3):535–574, 2019. doi: 10.1257/jel.20181020.
- [12] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30, pages 1024–1034, 2017.
- [13] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, 2021. doi: 10.1145/3457607.
- [14] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 652–660, 2017.
- [15] Mu Li, David G. Andersen, Jun Woo Park, Alexander J. Smola, Amr Ahmed, Vanja Josifovski, James Long, Eugene J. Shekita, and Bor-Yiing Su. Scaling distributed machine learning with the parameter server. In *Proceedings of the 11th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, pages 583–598, 2019.
- [16] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Lariviere, Alina Beygelzimer, Florence d’Alche Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research. In *ICML Workshop on Reproducibility in Machine Learning*, 2020.
- [17] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019. doi: 10.1056/NEJMra1814259.

- [18] George F. Jenks. The data model concept in statistical mapping. *International Yearbook of Cartography*, 7:186–190, 1967.
- [19] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- [20] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016. doi: 10.1145/2939672.2939785.
- [21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017.
- [23] David Ben-Shimon, Alexander Tsikinovsky, Michael Friedmann, Bracha Shapira, Lior Rokach, and Johannes Hoerle. Recsys 2015 challenge and the yahoo! music dataset. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 355–356, 2015. doi: 10.1145/2792838.2798723.
- [24] Yue Zhang, Wei Li, and Kai Chen. Conversion prediction from session data: A benchmark study. In *Proceedings of the 15th ACM Conference on Recommender Systems*, pages 123–132, 2021.
- [25] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. Neural attentive session-based recommendation. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1125–1134, 2020.