Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

Retentive Relevance: Capturing Long-Term User Value in Recommendation Systems

Saeideh Bakhshi, Phuong Mai Nguyen, Robert Schiller, Tiantian Xu, Pawan Kodandapani, Andrew Levine, Cayman Simpson, Qifan Wang

Meta

bakhshi@meta.com, pmnguyen@meta.com, rschiller@meta.com, tiantianx@meta.com, pawansk@meta.com, andrewlevine@meta.com, cayman@meta.com, wqfcr@meta.com

Abstract

Recommendation systems have traditionally relied on short-term engagement signals, such as clicks and likes, to personalize content. However, these signals are often noisy, sparse, and insufficient for capturing whether a recommendation supports future return to the platform. We introduce Retentive Relevance, a novel content-level survey-based feedback measure that directly assesses users' intent to return to the platform for similar content. Unlike other survey measures that focus on immediate satisfaction, Retentive Relevance targets forward-looking behavioral intentions and provides a stronger predictor of next-day retention. We validate Retentive Relevance using psychometric analyses suited to our single-item measures, establishing convergent, discriminant, and behavioral validity. Through large-scale offline modeling, we show that Retentive Relevance significantly outperforms both engagement signals and other survey measures in predicting next-day retention, especially for users with limited historical engagement. We develop a production-ready proxy model that integrates Retentive Relevance into the final stage of a multi-stage ranking system on a social media platform. Calibrated score adjustments based on this model yield improvements in engagement, retention, and content quality during a 14-day A/B experiment. This work links content-level user perceptions to short-horizon retention outcomes in production systems. We offer a scalable, user-centered approach with implications for responsible AI development.

Introduction

Recommendation systems are the backbone of modern digital platforms, guiding users through vast content libraries every day (Ricci, Rokach, and Shapira 2022). Yet, the core challenge remains: how do we transform sparse, noisy engagement signals into reliable predictions of user preferences and future return behavior (Herlocker et al. 2004)? Most large recommendation systems rely on user engagement signals such as clicks, likes, comments and dwell time, operating under the assumption that these actions accurately reflect user interests and will result in future engagement and retention (Hu, Koren, and Volinsky 2008; Koren, Bell, and Volinsky 2009). Yet, this engagement-centric approach is fundamentally limited. Large-scale studies reveal that users who interact with content do not always want more of the same (Sharma and Cosley 2013a; Hasan et al. 2024), and engagement signals are systematically biased toward popular items, often missing users' latent interests (Abdollahpouri, Burke, and Mobasher 2017; Raza et al. 2024). This disconnect is especially problematic when optimizing for retention, as short-term engagement frequently fails to predict sustained platform usage (Jannach et al. 2015; Gomez-Uribe and Hunt 2015; Xue et al. 2025).

To address these limitations, survey-based feedback has emerged as a promising alternative, offering direct insights into user preferences (McNee, Riedl, and Konstan 2006; Pu, Chen, and Hu 2011; Hasan et al. 2024; Lv et al. 2025). However, existing survey measures focus on capturing the immediate value of the recommendation for the user, lacking the forward-looking perspective required to optimize for retention. These retrospective measures, while informative, do not capture the behavioral intentions that drive users to return to platforms in the near term.

To bridge this gap, we introduce **Retentive Relevance**, a novel content-level survey measure designed to capture users' *intent to return* for similar content. By asking users immediately after a recommendation, "How likely or unlikely are you to return to [platform] to view more posts like this?", Retentive Relevance directly measures the value of a recommendation as it relates to users' intent to return, while maintaining the clarity and interpretability of survey-based self-reported feedback.

Our comprehensive evaluation—encompassing offline analyses, large-scale production deployments, and A/B experiments—demonstrates that *Retentive Relevance* consistently surpasses both traditional engagement signals and alternative survey measures in predicting next-day retention. In a 14-day online experiment, this leads to improved user retention, increased engagement, and measurable gains in content quality and integrity metrics. Notably, our approach is particularly effective for low-signal users, where conventional metrics are sparse or unreliable. The key contributions of this paper are:

- **Novel survey construct with predictive validity for retention:** Retentive Relevance is the first content-level survey measure empirically validated to predict next-day user retention in recommendation systems.
- **Complete operational framework:** We present an end-to-end methodology and framework— from survey de-

sign to production model deployment— for designing, validating, and operationalizing Retentive Relevance at scale in recommendation systems.

- **Large-scale experimental validation:** We provide robust evidence from live A/B experiments deployed in a large social media platform demonstrating that Retentive Relevance drives significant improvements in user retention, engagement, and content quality.
- **Theoretical and practical insights:** We highlight the superior predictive power of forward-looking behavioral intent, with implications for the design of more user-centered and responsible AI systems.

The remainder of this paper is organized as follows. Section 2 reviews related work and situates our contribution within the literature. Section 3 details our survey design, data collection, and bias correction methodology. Section 4 summarized the findings on the relationships between Retentive Relevance and other survey measures and engagement signals. Section 5 presents our offline retention modeling results and compares predictive performance of Retentive Relevance with other alternative survey measures. Section 6 covers our approach to building a proxy based on survey, integrating into ranking production system and results of our online A/B testing. Section 7 discusses implications and future directions.

Related Work

We structure our review of related work around four key areas: user feedback in recommendation systems, survey-based feedback mechanisms, retention prediction and long-term value, and methods for cold-start and low-signal users. This structure clarifies the landscape and highlights how our work advances the field.

User Feedback in Recommendation Systems

Recommendation systems utilize both implicit feedback (e.g., clicks, dwell time (Rendle et al. 2009)) and explicit feedback (e.g., ratings, thumbs up/down (Resnick et al. 1994; Adomavicius and Kwon 2012)). While implicit signals are abundant, they are often noisy and biased (Marlin et al. 2007; Joachims et al. 2005). Explicit feedback provides more direct signals but is less frequent and can be affected by response and popularity bias (Konstan and Riedl 2012). Most prior work focuses on immediate reactions to content, limiting the ability to predict long-term engagement (Chen, Zhang, and Zhou 2019; Zhang et al. 2014).

Survey-Based Feedback Mechanisms

Within explicit feedback, survey-based methods have gained prominence for their ability to directly capture user satisfaction and interest (Knijnenburg et al. 2012; Raza et al. 2024; Hasan et al. 2024). Surveys can complement behavioral metrics, address data sparsity, and improve model explainability (Lv et al. 2025; Covington, Adams, and Sargin 2016; Butmeh and Abu-Issa 2024). Research in this area explores optimal survey design, question framing, and timing (Kumar and Tomkins 2005; Liu, Dolan, and Pedersen 2010). However, most surveys are retrospective, evaluating the current consumption value of content rather than capturing forward-looking intent, a gap our work aims to address.

Retention Prediction and Long-Term Effectiveness

As digital platforms increasingly prioritize sustainable growth, accurately predicting user retention has emerged as a central challenge (Jhawar, Dharwadker et al. 2023; Sun et al. 2022). Recent research has introduced a range of advanced techniques to address this problem. Reinforcement learning frameworks have been developed to optimize cumulative long-term rewards (Zheng et al. 2018), while causal inference methods help disentangle the effects of specific content on user retention (Schnabel et al. 2016). Graph neural networks further enable the modeling of complex user-item-time interactions (Wu et al. 2019). In addition, multi-task and sequential modeling approaches have been proposed to balance short- and long-term objectives (Wang et al. 2018; Li et al. 2017). Adaptive retention optimization frameworks (Xue et al. 2025) and generative flow networks (Liu et al. 2024a,b) have demonstrated significant improvements in next-day return prediction and overall engagement, with large-scale deployments validating the practical impact of retention-focused systems (Cai et al. 2023). However, despite these advances, most existing methods continue to rely on noisy engagement signals and often lack explainable, recommendation-level approaches that can bridge the temporal gap between immediate user actions and future behavior.

Cold-Start and Low-Signal Users

The cold-start problem remains a fundamental challenge in personalization, particularly for new users or those with limited interaction history (Schein et al. 2002). Collaborative filtering methods are especially vulnerable to data sparsity, while content-based approaches may fail to capture valuable collaborative signals (Burke 2002). Hybrid models attempt to mitigate these issues by integrating multiple signal types, but they often still rely heavily on noisy implicit feedback (Adomavicius and Tuzhilin 2005). Recent advances have explored meta-learning, few-shot, and transfer learning techniques to address cold-start and low-signal scenarios (Vartak et al. 2017; Lee et al. 2019; Elkahky, Song, and He 2015). Additionally, large language models and graph-based methods have shown promise in extracting richer representations from auxiliary data (Zhang et al. 2025; Li et al. 2024). However, most focus on auxiliary data, more prone to accuracy issues, rather than capturing ground truth preferences via direct user feedback. Survey-based methods offer a distinct advantage in cold-start contexts by enabling the immediate collection of explicit user preferences, even in the absence of substantial behavioral history.

Our Contribution

This paper advances the field at the intersection of survey-based ground truth signal collection, retention prediction, and production-scale integration with recommendation systems. We introduce an end-to-end framework that is rigorously validated through large-scale offline analyses and

Name (Construct)	Survey Question	Sample size
Retentive Relevance (Likelihood to return)	How likely or unlikely are you to return to [Platform] to view more posts like this? Very likely, Likely, Neither likely nor unlikely, Unlikely, Very unlikely	$N = 63{,}708$
Interest Matching (Interest relevance)	To what extent does this video match your interests? A great deal, A lot, A moderate amount, A little, Not at all	$N = 58{,}872$
Worth Your Time (Recommendation value)	Was this video worth your time? Completely, Mostly, Somewhat, Barely, Not at all	$N = 76{,}263$

Table 1: "Retentive Relevance" was compared with two other survey measures. Each survey was administered under equal conditions but separately. Data collection occurred between December 2024 and January 2025 across 18 countries on a large social media platform targeted to a personalized video recommendation feed.

live online experiments. Our approach enables the direct integration of forward-looking user intent into algorithmic optimization, providing a scalable and interpretable signal for improving user retention. Beyond technical impact, our framework offers practical implications for broader responsible AI systems, supporting more user-aligned algorithmic systems and sustainable platform growth.

Survey Implementation, Data Collection and Bias Correction

In this section we discuss the theoretical foundation for this work and the approach for developing the survey instrument, validating it, collecting data and correcting bias.

Theoretical Foundation

We designed Retentive Relevance to capture users' forward-looking intentions to return to a recommendation platform based on the value they perceive in the content. Unlike other survey-based measures that focus on immediate value or interest relevance (See Table 1), Retentive Relevance specifically targets the antecedents of retention behavior. The item design is informed by the Theory of Planned Behavior (Ajzen 1991), which posits that behavioral intentions are strong predictors of actual behavior, as well as established research on behavioral intention measurement (Fishbein and Ajzen 1975). The key theoretical distinction between Retentive Relevance and other constructs lies in its temporal orientation and behavioral specificity. For example, Interest Matching (see Table 1) captures cognitive alignment between content and user preferences, while Worth-Your-Time assesses retrospective value. In contrast, Retentive Relevance explicitly probes the likelihood of future behavior, aligning more closely with the retention outcomes we aim to predict.

Survey Instrument Development

Following best practices in survey development (Tourangeau, Rips, and Rasinski 2000; Willis 2005; Groves et al. 2009), we employed a theory-driven, backwards-design approach. The survey item was formulated as: *"How likely or unlikely are you to return to [platform] to view more posts like this?"* where [platform] refers to the large-scale social media app where the survey was conducted. Responses were collected on a balanced 5-point Likert scale ranging from Very unlikely (1) to Very likely (5), with a neutral midpoint. The question wording was carefully crafted to specify the behavioral target ("return to [platform]"), clarify content specificity ("posts like this"), and capture likelihood rather than certainty, acknowledging the inherent uncertainty in predicting future behavior.

Construct Validation Protocol

To establish content validity (Brown 2015) and ensure comprehension across diverse user populations, we conducted cognitive testing following standardized protocols (Willis 2005; Tourangeau, Rips, and Rasinski 2000). Literature suggests that 5–12 participants are sufficient to identify most comprehension issues (Willis 2005). We recruited $N = 8$ participants from the United States, stratified by gender (50% female), age (18–24: 25%, 25–40: 50%, 41–65: 25%), and platform usage (active vs. infrequent users: 50%/50%). Each think-aloud session lasted approximately 30 minutes. Using standardized cognitive interviewing methods (Tourangeau, Rips, and Rasinski 2000), we systematically assessed four cognitive processes underlying survey response. We evaluated 1) *Comprehension* by asking participants "What does this question mean to you?" to assess understanding of the forward-looking, behavioral nature of the question. 2) *Retrieval processes* were examined through "What specific content were you thinking about?" to evaluate whether participants referenced the intended recommendation. 3) *Judgment formation* was assessed by asking "How did you decide on your rating?" to examine the decision-making process and influencing factors. Finally, 4) *Response mapping* was evaluated through "Was it easy to select from the provided options?" to assess the appropriateness of the scale and response burden. Results indicated consistent understanding of the Retentive Relevance construct, with an average inter-rater agreement of 87.5% on key comprehension items. Participants reliably distinguished Retentive Relevance from alternative measures (e.g., Interest

Matching and Worth Your Time) with 87.5% accuracy, as measured by the proportion who consistently identified the intended construct in comparison scenarios. Importantly, participants demonstrated clear conceptual differentiation between immediate content evaluation ("Was this good?") and future behavioral intention ("Will I come back for more like this?"), supporting the theoretical basis of our construct.

Survey Implementation

Surveys were implemented as a contextual overlay, appearing immediately after a video recommendation to minimize recall bias and maximize ecological validity. This timing ensures that users evaluate content while their experience and emotional response are still salient, reducing the cognitive burden and potential bias of retrospective evaluation (Tourangeau, Rips, and Rasinski 2000). The survey interface displayed a playable video thumbnail above the question (see example in Figure 1), allowing users to reference the content while responding. To mitigate response bias (Groves et al. 2009), we incorporated several design features including randomized response order to counteract order effects, balanced scale anchors to prevent directional bias, and a neutral midpoint to accommodate genuine ambivalence. Survey triggers were programmed to appear randomly across all video recommendations by feed position and regardless of user interaction (e.g. watched, engaged or skipped), ensuring unbiased sampling across the content valuation spectrum and preventing systematic exclusion of skipped content. Survey questions and response options were translated into users' local languages following established internationalization practices, with back-translation validation to ensure construct equivalence. The implementation was designed to ensure that data collection did not significantly disrupt the experience and always provided the option to skip the survey.

Data Collection

We collected survey responses $N = 63{,}708$ for Retentive Relevance, $N = 58{,}872$ for Interest Matching, and $N = 76{,}263$ for Worth Your Time under equivalent conditions and statistical treatment between December 2024 and January 2025 across 18 countries using stratified sampling by user engagement levels (active vs. less active users). For each survey response, we collected multi-level features at the user level (e.g. historical engagement, same-day engagement, next day engagement and demographics), content level (e.g. content topic, content age, overall content level engagements and popularity) and user-content level interactions (e.g. likes, comments, shares, watch time, skip, etc.).

Bias Correction

To address nonresponse bias, we implemented covariate balancing propensity scores following established practices (Schnabel et al. 2016; Joachims, Swaminathan, and Schnabel 2017). Our propensity score model incorporated user demographics (age cohorts, geographic regions, platform tenure), behavioral patterns (engagement and consumption levels), and platform features (tenure on platform).

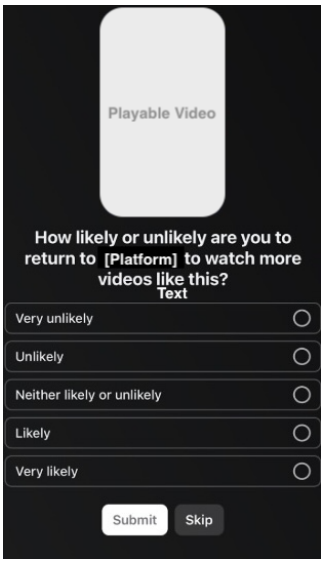

Figure 1: Schematic representation of the Retentive Relevance survey implementation. The interface maintains visual reference to the content being evaluated while capturing forward-looking behavioral intentions. "Platform" was replaced with the name of the social media app where the survey was deployed.

The Covariate Balancing Propensity Score (CBPS) optimization balances covariate distributions while estimating propensity scores:

$$\mathbb{E}[\pi(X_i)(1-\pi(X_i))X_i] = \frac{1}{n}\sum_{i=1}^{n}(Z_i-\pi(X_i))X_i = 0 \quad (1)$$

where $\pi(X_i)$ represents the propensity to respond to surveys and Z_i indicates survey completion. Post-weighting evaluation achieved standardized mean differences $|SMD| < 0.1$ across all covariates (Austin 2009), with trimming applied for extreme propensity scores following established practices (Crump et al. 2009).

Retentive Relevance vs. Alternative Surveys and Engagement Signals

To ensure that Retentive Relevance both captures the value of recommendations and remains distinct from other survey and engagement measures, we rigorously validated its psychometric properties —specifically, its convergent and discriminant validity— using established principles.

Convergent Validity and Relationships with Other Survey Measures

Convergent validity assesses whether measures that are theoretically related exhibit strong positive correlations, while still maintaining distinct conceptual boundaries (Campbell and Fiske 1959; Nunnally and Bernstein 1994). Following established validation protocols (Cohen 1988; Nunnally and Bernstein 1994), we evaluated convergent validity by

analyzing correlations between user-level survey responses within similar content types. To enable meaningful cross-sample comparisons, we first computed each user's mean response for a given measure within a content category, and then correlated those user-level category means across measures. The resulting correlations revealed significant positive associations among all measures, providing evidence for convergent validity. Notably, Retentive Relevance showed substantial correlations with Worth Your Time ($r = 0.63$, $p < 0.001$, 95% CI $[0.61, 0.65]$) and Interest Matching ($r = 0.58$, $p < 0.001$, 95% CI $[0.56, 0.60]$). These values fall within the optimal range for convergent validity (Cohen 1988), indicating meaningful conceptual overlap while remaining sufficiently below the threshold ($r < 0.85$) that would suggest redundancy (Kline 2015).

Discriminant Validity: What Makes Retentive Relevance Distinct

Discriminant validity requires that measures of theoretically distinct constructs display different response patterns across varied contexts (Campbell and Fiske 1959). We assessed this by examining how our survey measures differentiated between types of recommendation value across content categories. Our analysis revealed clear contextual differences in the relationships between measures. For content with immediate utilitarian or emotional value (e.g., motivation, learning, DIY), Retentive Relevance correlated more strongly with Worth Your Time (mean r = 0.69) than with Interest Matching (mean $r = 0.55$). In contrast, for interest-driven content (e.g., celebrities, technology, fashion), Retentive Relevance was more closely aligned with Interest Matching (mean $r = 0.65$) than with Worth Your Time (mean $r = 0.51$). This pattern suggests that Retentive Relevance adapts to different content contexts, capturing a broader spectrum of recommendation value. We further validated these differences using Fisher's z-transformation to compare correlation coefficients across content types. All observed differences were statistically significant ($z > 2.58$, $p < 0.01$), confirming that the measures respond systematically differently to distinct content topics.

Orthogonality to Existing Engagement Signals

To establish Retentive Relevance as a valuable and actionable signal for recommendation systems, it is crucial to demonstrate that it provides incremental information beyond what is captured by traditional engagement signals. We quantified the dependence between Retentive Relevance and standard engagement signals using mutual information, which measures how much knowing one variable reduces uncertainty about the other. As shown in Figure 2, the heatmap of mutual information coefficients reveals that Retentive Relevance consistently exhibits low mutual information with all traditional engagement signals ($MI < 0.15$ for all signals), indicating substantial independence from behavioral indicators. This finding suggests that user-stated intentions, as measured by Retentive Relevance, provide distinct and complementary information that is not fully captured by observed engagement alone, while also reflecting the fact that survey and behavioral measures arise from different measurement modalities.

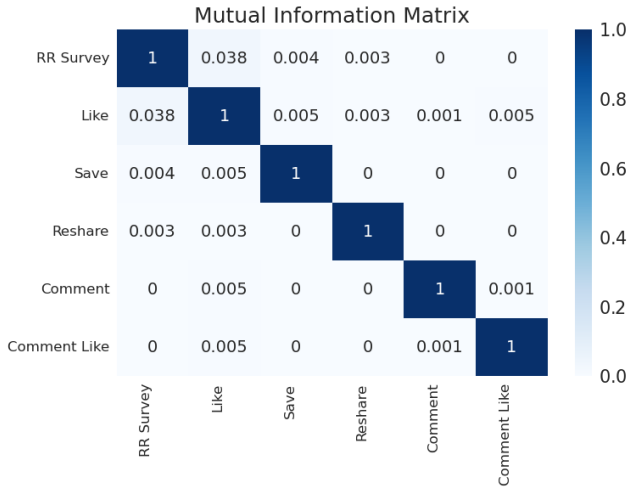

Figure 2: Heatmap of Mutual Information Matrix shows that Retentive Relevance captures information about recommendation that is distinct from engagement signals.

Predictive Performance of Survey-Based Signals in Retention Models

Having established that Retentive Relevance is a valid measure of personalized recommendation quality-demonstrating convergent validity while remaining distinct from other survey and engagement measures—we now evaluate its predictive power for next-day retention behavior. This analysis is designed to establish the behavioral validity of Retentive Relevance by comparing its predictive performance against alternative survey measures.

Modeling Approach

We formulate next-day retention prediction as a binary classification problem to assess the incremental value of survey responses. For each user i, the retention outcome $y_i \in \{0, 1\}$ indicates whether the user returns the following day, where $y_i = 1$ represents retention defined as video recommendation views exceeding the 5th percentile threshold of active user distributions. This operationalization distinguishes genuine retention behavior from accidental or minimal platform engagement.

We construct feature vectors $\mathbf{x}_i \in \mathbb{R}^d$ that capture multiple dimensions of user behavior and context. The feature vector is composed of five distinct components: $\mathbf{x}_i = [\mathbf{h}_i, \mathbf{r}_i, \mathbf{u}_i, \mathbf{c}_i, \mathbf{d}_i]$, where $\mathbf{h}_i$ represents historical engagement features aggregated over 28 days, $\mathbf{r}_i$ captures real-time signals including same-day activity patterns, $\mathbf{u}_i$ encompasses user-content interactions through both explicit and implicit feedback, $\mathbf{c}_i$ includes content metadata such as topic classification and creator characteristics, and $\mathbf{d}_i$ provides demographic and usage controls including age cohort and platform tenure.

Model	Overall Sample		Low-Signal Users	
	Accuracy (%)	ROC AUC	Accuracy (%)	ROC AUC
Baseline (No Survey)	78.0 ± 0.3	0.830 ± 0.005	73.0 ± 1.3	0.630 ± 0.013
+ Retentive Relevance	**83.0 ± 0.3*** **	**0.860 ± 0.005*** **	**76.0 ± 1.5*** **	**0.700 ± 0.025*** **
+ Worth Your Time	78.0 ± 0.4	0.828 ± 0.006	73.2 ± 1.4	0.632 ± 0.015
+ Interest Matching	78.2 ± 0.3	0.838 ± 0.005	73.1 ± 1.3	0.635 ± 0.014

Table 2: Predictive performance for next-day retention models shows that Retentive Relevance yields substantial and statistically significant gains in both accuracy and ROC AUC, while models with alternative survey measures do not show any statistically significant improvements. Results show mean ± 95% CI from stratified 10-fold cross-validation. $***p < 0.001$ compared to baseline via paired t-test. Bold indicates best performance for each metric.

We employ XGBoost gradient boosting classifiers optimized for log-loss, leveraging their robust performance with heterogeneous features and built-in regularization capabilities. The model prediction is formulated as:

$$\hat{y}_i = \sigma\left(\sum_{k=1}^{K} f_k(\mathbf{x}_i)\right) \tag{2}$$

where f_k represents the k-th tree in the ensemble, K denotes the total number of trees, and $\sigma(\cdot)$ is the sigmoid function mapping ensemble outputs to probability space. To assess the incremental value of each survey measure $s \in \{\text{RR}, \text{WYT}, \text{IM}\}$ (Retentive Relevance, Worth Your Time, Interest Matching), we construct paired model comparisons:

$$M_{\text{baseline}}: \quad P(y_i = 1|\mathbf{x}_i) \tag{3}$$

$$M_{\text{augmented}}: \quad P(y_i = 1|\mathbf{x}_i, s_i) \tag{4}$$

where s_i represents the survey response for user i. This paired design enables direct quantification of survey signal contributions while controlling for all other predictive factors.

We employ 10-fold cross-validation to maintain outcome class proportions across folds. For each fold $j \in \{1, \ldots, 10\}$, we compute performance metrics $\mathcal{M}_j$ including accuracy and ROC AUC for both baseline and augmented models. The incremental predictive value is quantified as the mean performance difference across folds:

$$\Delta\mathcal{M} = \frac{1}{10}\sum_{j=1}^{10}\left(\mathcal{M}_j^{\text{augmented}} - \mathcal{M}_j^{\text{baseline}}\right) \tag{5}$$

Statistical significance is assessed using paired t-tests across folds, with effect sizes calculated using Cohen's d. Bootstrap confidence intervals with 1000 iterations provide robust uncertainty estimates for performance improvements.

Predictive Performance Results

Table 2 presents the cross-validated performance results for next-day retention prediction with and without the survey measures. The results demonstrate that Retentive Relevance provides substantial and statistically significant improvements in both accuracy and ROC AUC. For the overall sample, incorporating Retentive Relevance into the prediction model increased accuracy by 5.0 percentage points (from 78.0% to 83.0%) and ROC AUC by 0.030 points (from 0.830 to 0.860), with significant effect sizes (Cohen's d = 2.1, $p < 0.001$).

The predictive gains were more pronounced for low-signal users, i.e. those with limited historical engagement data. For this user segment, Retentive Relevance increased accuracy by 3.0 percentage points and ROC AUC by 0.070 points (Cohen's d = 3.2, $p < 0.001$). The magnitude of these gains is particularly meaningful in large-scale recommendation systems, where even modest percentage increases can translate to additional retained users, especially considering these effects result from a single recommendation interaction. In contrast, neither Worth Your Time nor Interest Matching surveys provided significant predictive value for next-day retention, underscoring that Retentive Relevance captures unique behavioral intentions specifically relevant to retention decisions, rather than general content satisfaction or interest alignment captured by existing survey measures.

Feature Importance and Model Interpretation

We conducted feature importance analysis using SHAP (SHapley Additive exPlanations) values (Lundberg and Lee 2017), quantifying each feature's marginal contribution to individual predictions, expressed as percentage point changes in predicted retention probability(See Figure 3).

For the general population, "Unlikely" Retentive Relevance responses emerge as the second most important negative predictor (-2.1 pp), ranking immediately after same-day engagement controls. This substantial negative impact validates the behavioral connection between stated intent and actual retention outcomes, demonstrating that users who express low likelihood of returning indeed exhibit lower likelihood of returning for video views the next day. The effect becomes dramatically amplified for low-signal users, where "Very Likely" Retentive Relevance responses constitute the strongest positive predictor after controlling for same-day activity (+8.3 pp). This effect size substantially exceeds any traditional engagement factor, including likes, shares and comments. The magnitude of this impact underscores the particular value of direct intent measurement for users where behavioral signals are limited or unreliable. Across both user populations, Retentive Relevance responses consistently demonstrate superior predictive importance compared to traditional engagement measures. This disparity in-

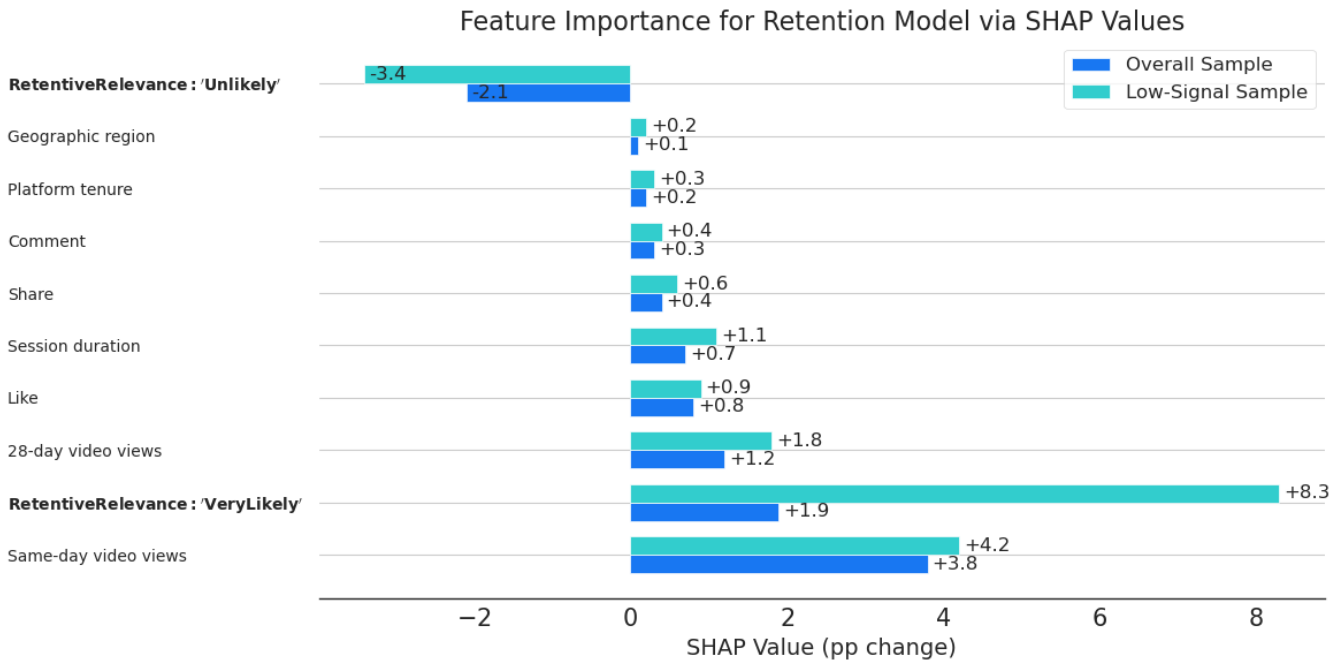

Figure 3: Feature importance analysis via SHAP values shows that Retentive Relevance significantly improves retention prediction—especially for low-signal users and with survey signals proving more predictive than engagement signals.

dicates that direct user intent signals provide substantially more predictive information than preferences inferred from behavioral observation alone.

These results establish that Retentive Relevance provides both statistically significant and practically meaningful improvements in retention prediction, with effect sizes that justify the implementation costs of survey-based feedback collection in production recommendation systems. The superior performance compared to established survey measures demonstrates that Retentive Relevance captures unique aspects of user experience specifically relevant to retention behavior, establishing its criterion validity as a forward-looking measure of user intent.

Production Integration and Online Evaluation

Having established the predictive validity of Retentive Relevance through comprehensive offline analysis, we now describe the end-to-end process of operationalizing survey signals within large-scale production recommendation systems. Our framework consists of three key phases: (1) development of production-ready proxy models that translate survey insights into real-time predictions, (2) integration of these predictions into existing ranking infrastructure through calibrated score adjustments, and (3) validation through large-scale online experimentation.

Survey Signal Proxy Model

We formulate survey signal prediction as a binary classification problem to estimate user retention intent for unseen user-item pairs. Let $\mathcal{U}$ and $\mathcal{V}$ denote the sets of users and items, respectively. For any user-item pair $(u, v) \in \mathcal{U} \times \mathcal{V}$, we aim to predict the probability that user u would express positive retention intent for item v. Given the 5-point Likert scale survey responses, we adopt a binary classification framework where positive intent corresponds to "Likely" or "Very Likely" responses (ratings 4-5), and negative intent corresponds to "Unlikely" or "Very Unlikely" responses (ratings 1-2). Neutral responses (rating 3) are excluded from training, as their inclusion decreased discriminative performance by 2.3% AUC. The proxy model is formulated as a logistic regression classifier optimized for production deployment:

$$P(\text{RR}_{u,v} = 1|\mathbf{x}_{u,v}) = \sigma(\mathbf{w}^T\mathbf{x}_{u,v} + b) \tag{6}$$

where $\sigma(\cdot)$ is the sigmoid function, $\mathbf{w}$ represents learned weights, b is the bias term, and $\mathbf{x}_{u,v} \in \mathbb{R}^d$ is the feature vector for user-item pair (u, v). The feature vector incorporates multiple signal categories following established practices (Covington, Adams, and Sargin 2016; Cheng et al. 2016):

$$\mathbf{x}_{u,v} = [\mathbf{p}_{u,v}, \mathbf{e}_u, \mathbf{c}_v, \mathbf{i}_{u,v}, \mathbf{n}_{u,v}] \tag{7}$$

where $\mathbf{p}_{u,v}$ represents behavioral prediction scores including learned probabilities for engagement actions, $\mathbf{e}_u$ captures temporal engagement rate features, $\mathbf{c}_v$ includes content metadata, $\mathbf{i}_{u,v}$ represents user-content interaction patterns, and $\mathbf{n}_{u,v}$ encompasses negative feedback indicators. The model is trained to minimize regularized logistic loss:

$$\mathcal{L}(\mathbf{w}, b) = -\frac{1}{N}\sum_{i=1}^{N}\Big[y_i \log p_i + (1-y_i)\log(1-p_i)\Big] + \lambda\|\mathbf{w}\|_2^2 \tag{8}$$

where N is the number of training samples, $y_i \in \{0, 1\}$ is the binary survey label, $p_i = P(\text{RR}_{u_i,v_i} = 1|\mathbf{x}_{u_i,v_i})$, and λ is the L2 regularization parameter.

Ranking Integration Architecture

Survey signal predictions are integrated into the final ranking stage of our multi-stage recommendation system on a large social media platform serving video recommendations. Let $\text{score}_{\text{base}}(u, v)$ denote the baseline ranking score for user u and item v. The survey-augmented ranking score is computed as:

$$\text{score}_{\text{final}}(u, v) = \text{score}_{\text{base}}(u, v) + \text{boost}(u, v) + \text{demote}(u, v) \tag{9}$$

where boost and demotion factors are defined as:

$$\text{boost}(u, v) = \alpha \cdot \mathbb{I}[\hat{p}_{u,v} > \tau_{\text{boost}}] \tag{10}$$

$$\text{demote}(u, v) = -\beta \cdot \mathbb{I}[\hat{p}_{u,v} < \tau_{\text{demote}}] \cdot (\tau_{\text{demote}} - \hat{p}_{u,v}) \tag{11}$$

Here, $\hat{p}_{u,v}$ is the predicted retention intent probability, $\alpha > 0$ and $\beta \in (0, 1)$ are tunable parameters, and τ_{boost} and τ_{demote} are precision-calibrated thresholds with $\tau_{\text{demote}} < \tau_{\text{boost}}$.

Threshold calibration follows a data-driven approach optimizing for precision and coverage. The boost threshold τ_{boost} achieves 80% positive precision at $\hat{p}_{u,v} > 0.76$, ensuring only high-confidence positive predictions receive ranking boosts. The demotion threshold τ_{demote} targets 60% negative precision at $\hat{p}_{u,v} < 0.38$, balancing sensitivity and specificity. Items with predicted probabilities in the neutral zone $[\tau_{\text{demote}}, \tau_{\text{boost}}]$ receive no treatment, maintaining ranking stability for uncertain predictions.

Online Experimental Results

We conducted large-scale online A/B experiments on a major social media platform with personalized video recommendations. The experimental design follows established best practices for recommendation system evaluation, incorporating comprehensive statistical rigor and multiple validation approaches.

We evaluated system performance across three primary metric categories: (1) User retention measured by sessions per user, (2) Engagement activity measured by metrics such as communication activity, like rates, and skip rates; (3) Content quality and integrity, measured through prevalence of reported content, negative feedback indicators, and established metrics based on quality and integrity classifiers. All metrics were tracked continuously throughout the experiment window, with statistical significance assessed using two-sample t-tests and effect sizes calculated using Cohen's d. Bootstrap confidence intervals provided robust uncertainty estimates for observed differences.

Table 3 summarizes the statistically significant changes observed across key platform metrics during the 14-day experimental period. The results demonstrate consistent improvements across user engagement, retention, and content quality metrics.

Category	**Metric**	**Change (% Δ ± 95% CI)**
Retention	Sessions per User	+0.030 ± 0.026
Engagement	Communication Activity	+0.052 ± 0.039
	Like Rate	+0.169 ± 0.100
	Skip Rate	–0.188 ± 0.085
Content Quality	Reported Content	–1.36 ± 0.11
	Negative Feedback	–1.527 ± 0.095
	"Not Interested" Feedback	–2.6 ± 1.3
	Reports to Likes Ratio	–0.825 ± 0.075

Table 3: Results from online A/B testing show that integrating Retentive Relevance into ranking leads to significant improvements in retention, engagement, and content quality metrics.All changes reported as percentage point differences with 95% confidence intervals. Negative values indicate reductions; positive values indicate increases. All reported changes are statistically significant at $p < 0.05$.

The treatment group showed enhanced user interaction patterns, with communication activity increasing by 0.052 percentage points (±0.039), like rates by 0.169 percentage points (±0.100), and skip rates decreasing by 0.188 percentage points (±0.085). Note that users often show their lack of interest in content via skip. Most critically, sessions per user increased by 0.030 percentage points (±0.026), reflecting improved retention.

Additionally, Retentive Relevance integration yielded improvements in content quality metrics. Prevalence of reported content decreased by 1.36 percentage points (±0.11), negative feedback declined by 1.527 percentage points (±0.095), and "not interested" signals dropped by 2.6 percentage points (±1.3). These reductions demonstrate the system's enhanced ability to identify and demote low quality content while improving user experience. The results demonstrate that optimizing for Retentive Relevance creates natural alignment between improved user experience, platform growth as well as improved content quality.

Discussion and Implications

Building on our results, we now explore the broader implications, limitations, and future directions of Retentive Relevance for recommendation systems and AI applications.

User-Centered Paradigm: Shifting the Foundation

Our work establishes a user-centered paradigm for recommender systems by empirically validating the connection

between content-level user perceptions and retention outcomes, aligning with the growing emphasis on intent-based AI systems. We show that survey responses capturing users' future intent are stronger predictors of actual behavior than traditional survey or engagement signals. The Theory of Planned Behavior (Ajzen 1991) informs this item design and interpretation by motivating the focus on behavioral intention, though our analyses should be understood as predictive validation rather than a full test of that theory. This theoretical framing distinguishes our approach from content-based and collaborative filtering methods that rely heavily on past interactions (Xu et al. 2025). The unique, orthogonal predictive power of Retentive Relevance—distinct from existing engagement metrics—demonstrates that intent-based feedback reveals different aspects of user preferences, addressing the limitation that users who interact with content do not always want more of the same (Hasan et al. 2024; Sharma and Cosley 2013b). This insight empowers platforms to move beyond optimizing for short-term engagement and to incorporate a forward-looking signal that is informative for near-term user retention.

Practical Impact and Industry Applications

We present an end-to-end framework, from survey design to production deployment, validated through large-scale online A/B testing. This production-ready approach is broadly applicable to other AI systems beyond recommendations, wherever user feedback and intent can help optimize or calibrate complex models. We show that optimizing for user intent drives simultaneous improvements in platform retention, engagement, and content quality metrics. These results demonstrate that user-centered optimization can resolve longstanding trade-offs between growth and responsibility—a critical consideration as organizations scale AI across multiple departments and business processes (IBM 2024). The measured improvements in content integrity metrics provide empirical evidence that intent-based optimization creates natural alignment between user experience and platform growth and quality.

Implications for Responsible AI Systems

Incorporating user feedback directly into AI systems has significant implications for responsible AI development. In our approach users directly express their intent, making algorithmic decisions more interpretable compared to systems that infer preferences from opaque behavioral signals. By enabling users to express their intent and preferences, recommendation algorithms become more transparent, accountable, and aligned with individual values. Ultimately, user-centered feedback mechanisms represent a step toward building AI systems that are not only effective but also better aligned with users' stated interests and values.

Limitations and Future Directions

While Retentive Relevance demonstrates strong effectiveness, several limitations present opportunities for future research. First, the empirical horizon in this paper is limited to next-day retention modeling and a 14-day online experiment. Accordingly, our results support Retentive Relevance as a useful forward-looking signal for short-horizon return behavior, but they do not by themselves establish longer-run user value effects. Currently, our approach also captures the value of a single recommendation interaction, missing the broader context of user sessions and sequence of recommendations. Future work could explore session-level and experience survey designs that can be used directly as optimization objectives rather than as additive signals, potentially incorporating advances in sequential modeling and multi-task learning (Raza et al. 2024). Longitudinal tracking can further illuminate the evolution of user intent over time, addressing how preferences shift across different contexts and temporal patterns. Additionally, expanding the framework to cross-modal recommendations and other AI systems could broaden its applicability.

Conclusion

In this paper, we introduce Retentive Relevance, a survey-based measure that advances the evaluation of the recommendation system from retrospective satisfaction to forward-looking user intent. By directly capturing the likelihood of users returning, we show that Retentive Relevance outperforms both traditional engagement signals and alternative survey measures to predict the next-day return to the platform. Integrating Retentive Relevance into ranking and validating it through a 14-day online A/B test, we show that it provides valuable additional signal on user preferences and improves retention, engagement, and content quality in the short horizon we study. Retentive Relevance offers a scalable, model-agnostic approach that bridges user perception research and production for more user-centered AI personalization.

References

Abdollahpouri, H.; Burke, R.; and Mobasher, B. 2017. Controlling popularity bias in learning-to-rank recommendation. In *Proceedings of the 11th ACM Conference on Recommender Systems*, 42–46. ACM.

Adomavicius, G.; and Kwon, Y. 2012. Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Transactions on Knowledge and Data Engineering*, 24(5): 896–911.

Adomavicius, G.; and Tuzhilin, A. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6): 734–749.

Ajzen, I. 1991. The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2): 179–211.

Austin, P. C. 2009. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine*, 28(25): 3083–3107.

Brown, T. A. 2015. Confirmatory factor analysis for applied research.

Burke, R. 2002. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4): 331–370.

Butmeh, H.; and Abu-Issa, A. 2024. Hybrid attribute-based recommender system for personalized e-learning with emphasis on cold start problem. *Frontiers in Computer Science*, 6: 1404391.

Cai, Q.; Liu, S.; Wang, X.; Zuo, T.; Xie, W.; Yang, B.; Zheng, D.; Jiang, P.; and Gai, K. 2023. Reinforcing User Retention in a Billion Scale Short Video Recommender System. In *Proceedings of the ACM Web Conference 2023*.

Campbell, D. T.; and Fiske, D. W. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, 56(2): 81–105.

Chen, L.; Zhang, G.; and Zhou, E. 2019. On the relationship between recommendation diversity and user satisfaction. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, 223–228. ACM.

Cheng, H.-T.; Koc, L.; Harmsen, J.; Shaked, T.; Chandra, T.; Aradhye, H.; Anderson, G.; Corrado, G.; Chai, W.; Ispir, M.; et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, 7–10.

Cohen, J. 1988. Statistical power analysis for the behavioral sciences.

Covington, P.; Adams, J.; and Sargin, E. 2016. Deep neural networks for YouTube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, 191–198. ACM.

Crump, R. K.; Hotz, V. J.; Imbens, G. W.; and Mitnik, O. A. 2009. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1): 187–199.

Elkahky, A. M.; Song, Y.; and He, X. 2015. A multi-view deep learning approach for cross domain user modeling in recommendation systems. In *Proceedings of the 24th International Conference on World Wide Web*, 278–288. ACM.

Fishbein, M.; and Ajzen, I. 1975. Belief, attitude, intention, and behavior: An introduction to theory and research.

Gomez-Uribe, C. A.; and Hunt, N. 2015. The Netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems*, 6(4): 1–19.

Groves, R. M.; Fowler Jr, F. J.; Couper, M. P.; Lepkowski, J. M.; Singer, E.; and Tourangeau, R. 2009. *Survey methodology*. John Wiley & Sons.

Hasan, E.; Rahman, M.; Ding, C.; Huang, J. X.; and Raza, S. 2024. Review-based Recommender Systems: A Survey of Approaches, Challenges and Future Perspectives. *arXiv preprint arXiv:2405.05562*.

Herlocker, J. L.; Konstan, J. A.; Terveen, L. G.; and Riedl, J. T. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1): 5–53.

Hu, Y.; Koren, Y.; and Volinsky, C. 2008. Collaborative filtering for implicit feedback datasets. In *Proceedings of the 8th IEEE International Conference on Data Mining*, 263–272. IEEE.

IBM. 2024. How to scale AI in your organization. https://www.ibm.com/think/topics/ai-scaling.

Jannach, D.; Lerche, L.; Kamehkhosh, I.; and Jugovac, M. 2015. What recommenders recommend: an analysis of recommendation biases and possible countermeasures. *User Modeling and User-Adapted Interaction*, 25(5): 427–491.

Jhawar, M.; Dharwadker, A.; et al. 2023. Quantifying and Leveraging User Fatigue for Interventions in Recommender Systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1–11. ACM.

Joachims, T.; Granka, L.; Pan, B.; Hembrooke, H.; and Gay, G. 2005. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference*, 154–161. ACM.

Joachims, T.; Swaminathan, A.; and Schnabel, T. 2017. Unbiased learning-to-rank with biased feedback. In *Proceedings of the tenth ACM international conference on web search and data mining*, 781–789.

Kline, R. B. 2015. *Principles and practice of structural equation modeling*. New York: Guilford Publications, 4 edition.

Knijnenburg, B. P.; Willemsen, M. C.; Gantner, Z.; Soncu, H.; and Newell, C. 2012. Explaining the user experience of recommender systems. volume 22, 441–504. Springer.

Konstan, J. A.; and Riedl, J. 2012. Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction*, 22(1-2): 101–123.

Koren, Y.; Bell, R.; and Volinsky, C. 2009. Matrix factorization techniques for recommender systems. *Computer*, 42(8): 30–37.

Kumar, R.; and Tomkins, A. 2005. A model-based approach for learning to rank. In *Proceedings of the 14th International Conference on World Wide Web*, 128–137. ACM.

Lee, H.; Im, J.; Jang, S.; Cho, H.; and Chung, S. 2019. MeLU: Meta-learned user preference estimator for cold-start recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1073–1082. ACM.

Li, J.; Ren, P.; Chen, Z.; Ren, Z.; Lian, T.; and Ma, J. 2017. Neural attentive session-based recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1419–1428. ACM.

Li, P.; et al. 2024. A Survey on Deep Neural Networks in Collaborative Filtering Recommendation Systems. *arXiv preprint arXiv:2412.01378*.

Liu, J.; Dolan, P.; and Pedersen, E. R. 2010. Personalized news recommendation based on click behavior. In *Proceedings of the 15th International Conference on Intelligent User Interfaces*, 31–40. ACM.

Liu, Z.; Liu, S.; Yang, B.; Xue, Z.; Cai, Q.; Zhao, X.; Zhang, Z.; Hu, L.; Li, H.; and Jiang, P. 2024a. Modeling User Retention through Generative Flow Networks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Liu, Z.; Liu, S.; Zhang, Z.; Cai, Q.; Zhao, X.; Zhao, K.; Hu, L.; Jiang, P.; and Gai, K. 2024b. Sequential Recommendation for Optimizing Both Immediate Feedback and Long-

term Retention. In *Proceedings of the 47th International ACM SIGIR Conference*.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 4765–4774.

Lv, M.; Hogg, D.; Grubb, T.; Bassi, S.; Li, M.; Simpson, C.; and Rajagopalan, S. 2025. Improve the Personalization of Large-Scale Ranking Systems by Integrating User Survey Feedback. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713644.

Marlin, B.; Zemel, R. S.; Roweis, S.; and Slaney, M. 2007. Collaborative filtering and the missing at random assumption. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, 267–275.

McNee, S. M.; Riedl, J.; and Konstan, J. A. 2006. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI'06 Extended Abstracts on Human Factors in Computing Systems*, 1097–1101. ACM.

Nunnally, J. C.; and Bernstein, I. H. 1994. *Psychometric theory*. New York: McGraw-Hill, 3 edition.

Pu, P.; Chen, L.; and Hu, R. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the 5th ACM Conference on Recommender Systems*, 157–164. ACM.

Raza, S.; Rahman, M.; Kamawal, S.; Toroghi, A.; Raval, A.; Navah, F.; and Kazemeini, A. 2024. A Comprehensive Review of Recommender Systems: Transitioning from Theory to Practice. *arXiv preprint arXiv:2407.13699*.

Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, 452–461.

Resnick, P.; Iacovou, N.; Suchak, M.; Bergstrom, P.; and Riedl, J. 1994. GroupLens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, 175–186. ACM.

Ricci, F.; Rokach, L.; and Shapira, B. 2022. *Recommender Systems Handbook*. Springer, 3rd edition.

Schein, A. I.; Popescul, A.; Ungar, L. H.; and Pennock, D. M. 2002. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference*, 253–260. ACM.

Schnabel, T.; Swaminathan, A.; Singh, A.; Chandak, N.; and Joachims, T. 2016. Recommendations as treatments: Debiasing learning and evaluation. In *Proceedings of the 33rd International Conference on Machine Learning*, 1670–1679.

Sharma, A.; and Cosley, D. 2013a. Do social explanations work? Studying and modeling the effects of social explanations in recommender systems. In *Proceedings of the 22nd International Conference on World Wide Web*, 1133–1144. ACM.

Sharma, A.; and Cosley, D. 2013b. Do social explanations work? Studying and modeling the effects of social explanations in recommender systems. In *Proceedings of the 22nd International Conference on World Wide Web*, 1133–1144.

Sun, Y.; et al. 2022. Surrogate for Long-Term User Experience in Recommender Systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1–9. ACM.

Tourangeau, R.; Rips, L. J.; and Rasinski, K. 2000. *The Psychology of Survey Response*. Cambridge University Press.

Vartak, M.; Thiagarajan, A.; Miranda, C.; Bratman, J.; and Larochelle, H. 2017. A meta-learning perspective on cold-start recommendations for items. In *Advances in Neural Information Processing Systems*, 6904–6914.

Wang, H.; Zhang, F.; Xie, X.; and Guo, M. 2018. DKN: Deep knowledge-aware network for news recommendation. In *Proceedings of the 2018 World Wide Web Conference*, 1835–1844. ACM.

Willis, G. B. 2005. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Sage Publications.

Wu, S.; Tang, Y.; Zhu, Y.; Wang, L.; Xie, X.; and Tan, T. 2019. Session-based recommendation with graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 346–353.

Xu, X.; Xu, Z.; Yu, P.; and Wang, J. 2025. Enhancing User Intent for Recommendation Systems via Large Language Models. *arXiv preprint arXiv:2501.10871*.

Xue, Z.; Cai, Q.; Liu, S.; Yang, B.; Zuo, T.; Hu, L.; Jiang, P.; Gai, K.; and An, B. 2025. AURO: Reinforcement Learning for Adaptive User Retention Optimization in Recommender Systems. In *Proceedings of the ACM Web Conference 2025*.

Zhang, W.; Bei, Y.; Yang, L.; Zou, H. P.; Zhou, P.; Liu, A.; Li, Y.; Chen, H.; Wang, J.; Wang, Y.; et al. 2025. Cold-start recommendation towards the era of large language models (llms): A comprehensive survey and roadmap. *arXiv preprint arXiv:2501.01945*.

Zhang, Y.; Lai, G.; Zhang, M.; Zhang, Y.; Liu, Y.; and Ma, S. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th International ACM SIGIR Conference*, 83–92. ACM.

Zheng, G.; Zhang, F.; Zheng, Z.; Xiang, Y.; Yuan, N. J.; Xie, X.; and Li, Z. 2018. DRN: A deep reinforcement learning framework for news recommendation. In *Proceedings of the 2018 World Wide Web Conference*, 167–176. ACM.

Checklist

1. For most authors...

   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, we study a user-centered survey signal for recommender evaluation using privacy-protected, aggregated platform data; the survey is optional and designed to reduce bias (see Survey Implementation, Bias Correction, and Discussion).

(b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, the abstract/introduction claim an intent-to-return survey signal and we validate it via psychometrics, offline retention prediction, and online A/B tests.

(c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, and we describe an end-to-end approach (survey design + validation, bias correction, offline modeling, and production integration) matched to the claims (see Methods sections throughout).

(d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, we discuss population and context artifacts via multi-country sampling, low-signal vs overall analyses, and nonresponse-bias correction with propensity weighting (see Data Collection and Bias Correction).

(e) Did you describe the limitations of your work? Yes, we describe limitations and future directions (e.g., single-interaction focus, need for session-level/longitudinal extensions) in Discussion and Limitations/Future Directions.

(f) Did you discuss any potential negative societal impacts of your work? Yes, we discuss risks such as optimizing for retention at the expense of well-being, survey burden, and potential representation issues; we also report content-quality impacts in online tests (see Discussion and Online Experimental Results).

(g) Did you discuss any potential misuse of your work? We discuss potential biases with self selection bias potentially impacting recommendations and we discuss our approach to mitigate bias

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, and we mitigate risks via an opt-out survey design, randomized triggering, internationalization/back-translation, bias correction, interpretable calibration thresholds, and limiting reporting to aggregate results; data/code access is controlled (see Survey Implementation, Bias Correction, and Ranking Integration Architecture).

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes, and we have reviewed the ICWSM ethics guidelines and designed the study to minimize user risk, protect privacy, and avoid disclosure of sensitive platform details.

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? NA (the paper is primarily an empirical measurement and modeling study; we do not present formal hypothesis tests as the main contribution).

(b) Have you provided justifications for all theoretical results? NA (no theoretical results are presented as hypothesis-driven tests; empirical analyses are reported with appropriate uncertainty estimates).

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA (the work is framed around prior theory, e.g., Theory of Planned Behavior, but not as competing hypothesis tests).

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA (we focus on empirical validation and predictive comparisons rather than adjudicating alternative causal mechanisms).

(e) Did you address potential biases or limitations in your theoretical framework? NA (no separate theoretical framework is asserted beyond established behavioral-intent theory and standard recommender evaluation practice).

(f) Have you related your theoretical results to the existing literature in social science? NA (we situate results in prior literature in Related Work, but this checklist subsection is not applicable as hypothesis-testing theory development).

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA (implications are discussed, but not as policy-relevant theoretical hypothesis results).

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA (no formal proofs are included).

(b) Did you include complete proofs of all theoretical results? NA (no formal proofs are included).

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? No, because the underlying platform data, pipelines, and production systems are proprietary and cannot be shared; we provide detailed methodological descriptions and evaluation protocols in the paper.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? No, because some training specifics (e.g., complete feature definitions and internal hyperparameter choices) are not fully disclosed; we do specify the model families, label definitions, and validation approach (see Modeling Approach).

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes, and we report uncertainty via 95% confidence intervals and cross-validation summaries (e.g., Table 2; Table 3).

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No, because total compute and resource details (e.g., internal clusters) are not reported in the manuscript.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, and the evaluation triangulates psychometric validation, offline predictive performance (overall and low-signal users), and online A/B testing, matching the claims (see Sections on Validation, Retention Modeling, and Online Experimental Results).

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, and we operationalize fault tolerance via precision-calibrated thresholds for boosting/demotion (e.g., targeting 80% positive precision for boosts and 60% negative precision for demotions), which encodes the cost of false positives/negatives in ranking (see Ranking Integration Architecture).

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes, and we cite relevant prior work throughout Related Work and Methods (e.g., recommender evaluation, survey methodology, SHAP, and learning-to-rank literature).

(b) Did you mention the license of the assets? NA (we do not release or rely on externally licensed datasets/code; the data and implementation are internal/proprietary).

(c) Did you include any new assets in the supplemental material or as a URL? No, because we do not include new datasets or code artifacts due to proprietary constraints; we do include the survey questions/options and a schematic of the UI in the paper (Table 1 and Figure 1).

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, and survey participation was optional (users could skip), and the study follows the platform's standard consent and privacy processes for product research (see Survey Implementation and Data Collection).

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, and we avoid sharing personally identifiable information and report only aggregate analyses; the platform contains user-generated content, so content-quality and safety considerations are evaluated via reported-content and negative-feedback metrics in online tests (see Online Experimental Results).

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see **?**)? NA (we are not curating or releasing a new dataset).

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see **?**)? NA (we are not curating or releasing a new dataset).

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? Yes, and we include the exact survey items and response options (Table 1), and provide an interface schematic (Figure 1) and cognitive testing protocol description (Construct Validation Protocol).

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes, the study is designed to be minimal risk and privacy-preserving.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation?
answerNoNo, the survey data was collected as in-product intercepts, were optional and were not compensated.

(d) Did you discuss how data is stored, shared, and deidentified? No, the operational details on storage/sharing/deidentification is propriety and these processes are governed by company internal access controls and privacy standards.