

LLM-based User-Product Relevance Evaluation for Dynamic Product Advertising

Wen Zhang
Snap Inc.
Bellevue, WA, USA
wzhang7@snapchat.com

Xiao Bai
Snap Inc.
Palo Alto, CA, USA
xbai@snapchat.com

Hsiang-wei Chao
Snap Inc.
Seattle, WA, USA
hchao@snapchat.com

Mengfan Zhu
Snap Inc.
Santa Monica, CA, USA
mzhu2@snapchat.com

Tong Zhao
Snap Inc.
Bellevue, WA, USA
tzhao@snapchat.com

Mingxuan Ju
Snap Inc.
Bellevue, WA, USA
mju@snapchat.com

Xiaodong Liu
Snap Inc.
Bellevue, WA, USA
xliu9@snapchat.com

Congfei Zhang
Snap Inc.
Bellevue, WA, USA
czhang3@snapchat.com

Yunzhi Zhou
Snap Inc.
Palo Alto, CA, USA
yzhou10@snapchat.com

Yajun Wang
Snap Inc.
Palo Alto, CA, USA
ywang30@snapchat.com

Jinchao Li
Snap Inc.
Bellevue, WA, USA
jli18@snapchat.com

Yu Zhang
Snap Inc.
Palo Alto, CA, USA
yzhang3@snapchat.com

Abstract

Dynamic Product Advertising (DPA) retrieval systems are a core component of modern e-commerce advertising; however, their evaluation remains challenging. Conventional offline engagement metrics are limited by sparse, biased feedback, while online evaluation is costly and time-consuming. In this work, we propose a user-product relevance metric for offline evaluation of DPA retrieval systems. The metric first compresses the engagement history into a compact set of intent representatives via Semantic-ID-grouped, event-stratified segmentation, then assesses the relevance of a recommended product to each individual intent and aggregates them as a user-level relevance score, yielding a stable and interpretable measure of retrieval quality. To compute this metric at scale, we leverage large language models (LLMs) as semantic relevance judges, capturing nuanced functional and semantic relationships that heuristic methods often miss. We propose an agentic, human-aligned prompting framework that generates multi-category relevance guidelines and distills teacher-model judgments into an efficient student model for scalable evaluation. The resulting metric strongly correlates with offline and online business metrics, enabling faster and more reliable DPA retrieval iteration before A/B testing.

CCS Concepts

• Information systems → Relevance assessment.

Keywords

LLM Judge, DPA, Semantic ID, Agentic, Offline Evaluation, Product Relevance, E-Commerce

ACM Reference Format:

Wen Zhang, Xiao Bai, Hsiang-wei Chao, Mengfan Zhu, Tong Zhao, Mingxuan Ju, Xiaodong Liu, Congfei Zhang, Yunzhi Zhou, Yajun Wang, Jinchao Li, and Yu Zhang. 2026. LLM-based User-Product Relevance Evaluation for Dynamic Product Advertising. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26)*, November 07–11, 2026, Rome, Italy. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3799682.3840117>

1 Introduction

Modern E-commerce advertising platforms increasingly rely on high-quality Dynamic Product Ads (DPA) to deliver personalized user experiences at scale. Within these systems, the major task of retrieval stage is efficiently filtering a massive product catalog into a small subset of candidates based on diverse user interests and engagement histories. DPA retrieval systems typically rely on multiple algorithms to retrieve candidate products to ensure diversity. Although the top-line business objective is to maximize user engagement and revenue, the industry lacks a universal and reliable metric to compare the effectiveness of different retrieval algorithms. Common offline evaluation metrics such as the recall of future engagements [19] are limited by selection bias and data sparsity. Unlike in the Search domain, where explicit queries provide a clear benchmark for success, the advertising domains lack a well-defined framework for user-product relevance [1]. Consequently, machine learning practitioners often struggle to quantify whether a retrieval model is successfully surfacing relevant products or merely exploiting noisy, short-term behavioral correlations that may not lead to sustainable long-term growth.



This work is licensed under a Creative Commons Attribution 4.0 International License. *CIKM '26, Rome, Italy*

© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2539-5/2026/11
<https://doi.org/10.1145/3799682.3840117>

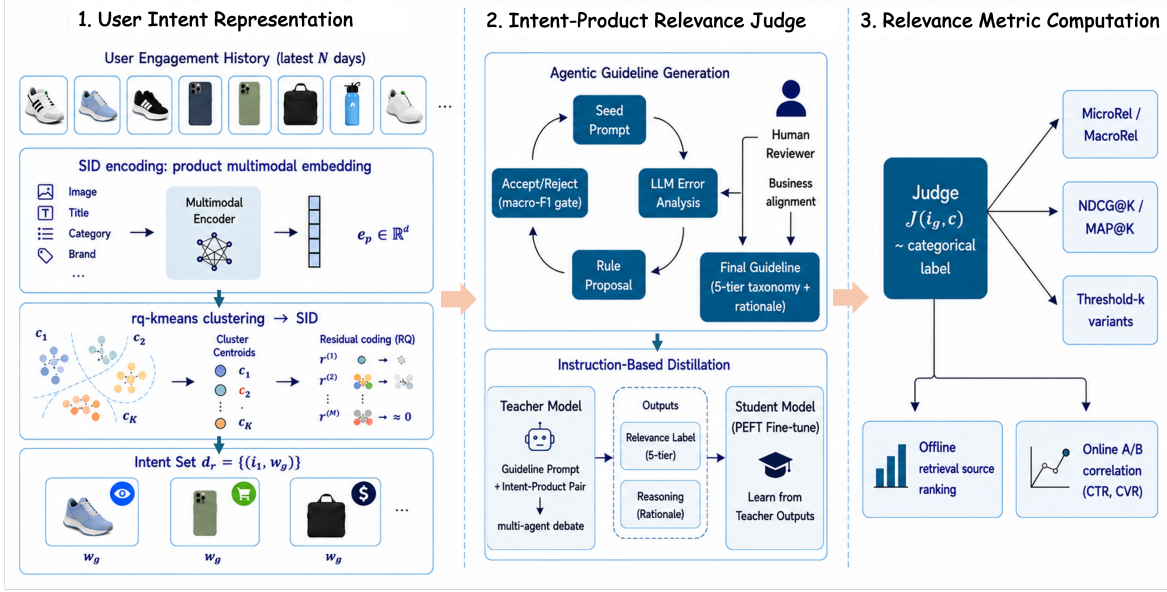


Figure 1: LLM-based User-Product Relevance Evaluation framework. (1) **User Intent Representation:** a multimodal encoder plus residual vector quantization assigns each engaged product a hierarchical semantic ID (SID); events sharing (SID, e) collapse into a single weighted intent representative, yielding the compact intent set d_r . (2) **Intent-Product Relevance Judge:** an agentic guideline aligned with human reviewers anchors a teacher, PEFT-distilled into a student model. (3) **Metric Computation:** MicroRel/MacroRel under threshold- k aggregation, validated against offline retrieval-source rankings and online CTR/CVR.

Defining and measuring relevance is the core obstacle. Prior work either uses explicit user studies that rate usefulness, trust, or satisfaction [9, 12], or implicit engagement logs of clicks, likes, shares, and reports [19]. User studies are expensive and slow [12]. Engagement is abundant but noisy: clicks are biased by presentation and trust effects [5], and engagement-optimized ranking amplifies attention-grabbing content users do not actually prefer when asked [6]. We instead use a content-based definition of relevance: **an LLM judges each retrieved product against intents inferred from the user’s engagement history, and intent-level scores are aggregated into a user-level metric.** This reduces dependency on exposure artifacts, covers users with sparse-feedback, and makes the relevance score interpretable at the attribute level.

Recent work uses LLMs as semantic relevance judges for search and recommendation, scoring content alignment directly rather than relying on engagement signals [2, 10, 14, 15]. Sun et al. [14] show that prompted LLMs are effective zero-shot re-rankers and can outperform strong supervised baselines; Choi et al. [2] report substantial agreement between LLM evaluators and human raters; and large-scale benchmarks find that LLM-generated relevance labels broadly track human judgment at scale [10]. These works target search-style relevance with an explicit query. In contrast, DPA retrieval system do not have search queries and typically require inferring user-level relevance signals from heterogeneous engagement histories.

We propose an offline evaluation framework that uses LLM judges to score user-product relevance for DPA retrieval. Two operational problems shape the design: the engagement history is heterogeneous and long, and frontier LLMs are too expensive to call

on every event-candidate pair at production volume. We address both with (i) an SID (Semantic ID [13])-grouped, event-stratified compression of engagement history into a small set of weighted intent representatives, and (ii) an agentic guideline-discovery loop together with instruction-guided teacher-student finetune for production serving. Our contributions are 3-folds as follows:

- **Query-free relevance metric.** We define a user-level offline evaluation metric for DPA retrieval that aggregates LLM-judged intent-candidate relevance under a threshold- k aggregation rule, without requiring an explicit query.
- **A Scalable LLM-judge Pipeline.** We propose an efficient LLM judge pipeline with 1) SID-grouped intent decomposition that reduces LLM calls by 30 $\times$, and 2) teacher-student distillation from an agentic, human-vetted guideline that replaces the frontier teacher with an efficient student at production cost.
- **Online validation.** The metric ranks six retrieval sources consistently with conversion precision/recall - how well the retrieved candidates aligns with user’s subsequence purchases - and is directionally aligned with online CTR/CVR deltas across four A/B tests; It thus provides a reliable view of retrieval system health and a low-cost signal for guiding offline model iteration.

2 LLM-Based User-Product Relevance Metric

The proposed relevance evaluation framework has three blocks (Figure 1): a compact *user intent representation* that segments engagement history into SID-grouped representatives (§2.1); an *intent-product relevance judge* built from an agentic-discovered guideline (§2.2.1) and a teacher-student distillation (§2.2.2); and a *metric-computation layer* (§2.3) that aggregates pairwise judgments into request- and user-level scores.

2.1 User Intent Representation via SID-based Segmentation

A user is represented by her engagement history $H_u = \{(p_t, e_t)\}_{t=1}^T$, where p_t is an engaged product and e_t is the event type (view, add-to-cart, purchase, etc.). Treating every event as a distinct intent is the most expressive representation but costly in production: a typical 30-day window holds hundreds of events, most of which are semantically redundant, inflating LLM cost and biasing the metric toward retrieval algorithms that re-surface near-duplicates of frequently-viewed items.

SID grouping. To select a compact, signal-rich representative set, we run a lightweight grouping step before the judge. Each product in the catalog is mapped to a hierarchical *Semantic ID* (SID) via k-means residual vector quantization (RQ-kmeans) over an in-house content-based multimodal product embedding [3], using L stages with 512-entry codebooks to provide sufficient granularity. Concretely, with $r_0 = \text{embed}(p)$ and $C_l \in \mathbb{R}^{512 \times d}$ the level- l codebook,

$$\text{SID}(p) = (c_1, \dots, c_L), \quad c_l = \arg \min_k \|r_{l-1} - C_l^{(k)}\|_2, \quad r_l = r_{l-1} - C_l^{(c_l)}, \quad (1)$$

so each stage encodes the residual left by the previous one and longer prefixes refine coarser ones. The length of the leading prefix is used as the grouping key to control granularity: a length-1 prefix groups products at roughly category-prototype resolution (e.g., all running shoes), while the full length- L prefix distinguishes near product-type-plus-attribute granularity (e.g., a specific brand-model-color combination). Two items sharing the chosen-length prefix are treated as content-equivalent; The SID lookup is precomputed and cached, so its runtime cost is negligible during LLM judge inference. With $L=3$ in production – a balance between cost (fewer LLM calls at shallow L) and granularity (longer prefixes preserve product-level distinctions) – we partition H_u by $(\text{SID}(p_t), e_t)$ and elect one representative i_g per group – the most recent product, with ties broken by event-type strength (e.g., *purchase* > *add-to-cart* > *others*). The resulting weighted intent set is

$$d_u = \{(i_g, w_g) : g \in \text{Groups}(H_u)\}, \quad w_g = \log(1 + |g|) \cdot \beta(e_g), \quad (2)$$

with $\beta(\text{purchase})=3$, $\beta(\text{add-to-cart})=2$, $\beta(\text{others})=1$. The $\log(1 + |g|)$ factor lets a heavily-engaged content-equivalence group dominate over a one-off view of an unrelated product without scaling linearly with view count; event-type stratification keeps a *low-intent* signal from overruling a same-SID *high-intent* signal.

2.2 Intent-Product Relevance Judge

Given the intent set d_r from §2.1, we need a judge $\mathcal{J}(i, c)$ that assigns a relevance label to each (intent representative i , candidate product c). Two issues shape the design. The taxonomy must capture DPA-specific product semantics without becoming so detailed that a frontier LLM applies it inconsistently; and running a frontier LLM as $\mathcal{J}$ at production volume is too expensive. We address the first with an agentic guideline-discovery loop followed by a human-alignment pass (§2.2.1), and the second with knowledge distillation into a small student (§2.2.2).

2.2.1 Agentic Guideline Generation. We build the evaluation guideline in two phases. Phase A is an *agentic loop* where an LLM judge iteratively refines its own decision rules from labeled examples;

Phase B is a *human alignment* pass where editors steer the resulting rule set toward DPA business priorities.

Phase A: agentic rule discovery. We start from a minimal *criteria-only* seed prompt – a single instruction, “*Label the candidate as Exact, Similar, Substitute, Complementary, or Irrelevant*” – with no per-class definitions, examples, or attribute hints. A frontier LLM (e.g., Gemini-2.5-Pro) labels a stratified calibration split using this seed prompt, scored by per-class F1 against the human labels. The refinement loop iterates (i) clustering mis-classified pairs by the (predicted, label) confusion and the judge’s free-form rationale; (ii) prompting a separate reasoner to propose up to three new sub-rules targeting the true-label class for the largest cluster; and (iii) provisionally appending each sub-rule, re-labeling the calibration set, and accepting iff the paired-bootstrap 95% CI on Δ accuracy excludes zero with no per-class F1 regression beyond 0.5 percentage points. Rejected sub-rules are logged as “previously tried”. A round ends on the first accepted sub-rule or after all proposals are rejected; the loop stops after N rounds or two consecutive empty rounds. We run multiple trials and retain only sub-rules that recur across trials, stabilizing the discovered set against proposer stochasticity. The candidate-rule pool surfaces seven of the twelve editorial-guideline reasons (attribute-level: type, brand, model/style, size, color, material, component-vs-whole); the five not surfaced – demographic gender/age, broad-Complementary/Irrelevant, asymmetric Complementary – are deferred to Phase B.

Phase B: human business-alignment. The agentic phase optimizes calibration-set agreement but is silent on downstream business goals. Human reviewers audit the Phase-A output against three criteria: coverage of business-critical mismatches (demographic gender/age, regulated categories, advertiser exclusions), editorial-standard compatibility, and reproducibility on a 200-pair business-stratified sample. Vetoed rules return to Phase A with an explanatory note. The most consequential intervention in our deployment promoted demographic mismatch from a *Substitute*-leaning signal to a hard *Irrelevant* override – a rule the loop repeatedly deprioritized because demographic-mismatch pairs are statistically rare in the calibration set. Phase B is load-bearing: demographic-override rules emerge from editorial review rather than from the agentic loop, because demographic-mismatch pairs are too rare in the calibration set to drive the loop’s statistical gate.

Final taxonomy and rationale format. The resulting guideline – the seed taxonomy refined by Phase-A sub-rules under the Phase-B business overrides – is what the rest of the paper refers to as “the guideline”. It instructs the LLM judge to assign each (intent, candidate) pair to one of five tiers and to emit a structured rationale by selecting from a controlled attribute vocabulary (*Product Type, Brand, Demographics, Component-vs-Whole*, etc.). The final taxonomy summarization:

- **Exact:** matches all specific attributes of the intent (type, brand, model, size, color).
- **Similar:** same product category, differs in brand, model, or color.
- **Substitute:** different type but functionally interchangeable.
- **Complementary:** different type, typically used together with the intent [4].

- **Irrelevant:** no functional or categorical link, or a critical demographic mismatch.

2.2.2 Teacher-Student Knowledge Distillation. The finalized guideline from §2.2.1 could in principle be deployed as an instruction prompt to a frontier LLM at serving time, but at production volume this is prohibitively expensive (Table 5). We instead use the guideline-prompted frontier LLM as a *teacher* that emits high-quality labels and rationales on a curated training corpus, then fine-tune a much smaller *student* LLM on those labels. The student is what we deploy as $\mathcal{J}$.

Distillation dataset construction. We construct a large distillation corpus that reflects real-world retrieval scenarios via a four-stage pipeline. **Stratified seed sampling:** intent representatives i_g are sampled stratified across user engagements and demographics (age, gender, engagement type, engagement frequency), ensuring broader and more balanced pattern coverage than a uniform random distribution. **Multi-source candidate generation:** for each seed, candidates are aggregated from three complementary streams – production retrieval logs, historical conversion-driven engagement signals, and curated high-value advertiser groups. **Hard-case mining:** we downsample to $N = 10$ candidates per seed using a multimodal embedding filter to intentionally surface semantically-similar-but-irrelevant negatives. **Labeling:** pairs are labeled by the teacher protocol above; rationale traces are retained as auxiliary supervision for the student.

Teacher labeling. The finalized guideline is loaded as the instruction prompt to a frontier LLM (Gemini-2.5-Pro), which annotates each seed–candidate pair in the distillation datasets with a relation label and an attribute-level rationale. GPT-5 runs in parallel under the same guideline as a verification agent; pairs on which the two agents agree at high self-reported confidence are accepted directly, and the disagreed cases are sent to human auditors for relabeling.

Student fine-tuning. We test both Gemini-2.5-Flash and Gemini-2.5-Flash-Lite as student backbones and apply parameter-efficient instruction fine-tuning (LoRA with adapter rank 16, learning-rate multiplier 2, 5 epochs). The student is trained for multi-label classification over the five-tier taxonomy and jointly emits an attribute-level rationale conditioned on the label, matching the teacher’s output format. At serving time, the student replaces both teacher LLMs in the inference path; the teacher is invoked only on guideline-update or distillation-refresh cycles.

2.3 Metric Computation

Given the intent set d_u from §2.1 and the pairwise judge $\mathcal{J}$ from §2.2, we compute a per-user relevance score $s(u)$ and then aggregate across users into a single retrieval-quality metric. The aggregation layer is decoupled from the upstream blocks so the same judgments can be re-aggregated under different rules without re-invoking the LLM. Algorithm 1 summarizes the procedure.

Per-user score. For each user u and candidate $c \in C_u$, the judge $\mathcal{J}$ emits a categorical label that we map to a per-pair score (binary $\{0, 1\}$ by default; alternatively a 5-grade gain $g \in \{0, 1, 2, 3, 4\}$ following [11]). The weighted intent-coverage $m_u(c) = \sum_{(i_g, w_g) \in d_u} w_g \cdot$

Algorithm 1: Offline User–Product Relevance Metric

Input: A group of users G (e.g., users routed to a specific retrieval source); per-user engagement history $H_u = \{(p_t, e_t)\}$ and candidate set C_u for each $u \in G$; SID truncation L ; event-type weights $\beta(\cdot)$; threshold k ; judge $\mathcal{J}(i, c)$.

Output: Per-user scores $\{s(u)\}_{u \in G}$; group-level summaries $\text{MicroRel}(G)$, $\text{MacroRel}(G)$.

// Step 1: per-user intent set + relevance score

foreach $u \in G$ **do**

$\mathcal{P}_u \leftarrow \text{Partition}(H_u, \text{key}=(\text{SID}_L(p_t), e_t))$;

$d_u \leftarrow \{(i_g, w_g) : g \in \mathcal{P}_u\}$, where

$i_g = \arg \max_{(p, e) \in g} \text{recency}(p)$, $w_g = \log(1 + |g|) \cdot \beta(e_g)$;

foreach $c \in C_u$ **do**

$m_u(c) \leftarrow \sum_{(i_g, w_g) \in d_u} w_g \cdot \mathcal{J}(i_g, c)$; // weighted intent-coverage

$v_u(c) \leftarrow \mathbb{I}[m_u(c) \geq k]$; // candidate-level binary relevance

$s(u) \leftarrow \frac{1}{|C_u|} \sum_{c \in C_u} v_u(c)$; // per-user score

// Step 2: group-level aggregation

$\text{MicroRel}(G) \leftarrow \frac{\sum_{u \in G} \sum_{c \in C_u} v_u(c)}{\sum_{u \in G} |C_u|}$; // (u, c) -equal weighting

$\text{MacroRel}(G) \leftarrow \frac{1}{|G|} \sum_{u \in G} s(u)$; // user-equal weighting

return $\{s(u)\}_{u \in G}$, $\text{MicroRel}(G)$, $\text{MacroRel}(G)$;

$\mathcal{J}(i_g, c)$ aggregates the per-pair signal across the intent representatives, and the threshold k converts it to a binary candidate-level relevance $v_u(c) = \mathbb{I}[m_u(c) \geq k]$. The per-user score $s(u) = \frac{1}{|C_u|} \sum_{c \in C_u} v_u(c)$ is the primary output of the metric layer.

User vs. group aggregation. The score $s(u)$ can be reported per user, or aggregated to any group G of users (retrieval source, demographic bucket, etc.). We use two complementary summaries: **MicroRel** $(G) = \sum_{u \in G} \sum_c v_u(c) / \sum_{u \in G} |C_u|$ weights each (u, c) pair equally; **MacroRel** $(G) = \frac{1}{|G|} \sum_{u \in G} s(u)$ weights each user equally and is therefore less sensitive to skewed traffic. In the rest of the paper we group by retrieval source; the same machinery applies to any other user grouping.

Threshold k as a strictness knob. We operate at $k = 7$ in production (justified in §4.4); the simpler $k = 1$ at-least-one rule is included as a permissive baseline since it equates a single weak intent match with broad intent coverage.

3 Online-Offline Retrieval Metrics Alignment

In this section, we benchmark the proposed LLM-based user-product relevance metric against conventional DPA retrieval offline metrics (e.g., **Conversion Precision/Recall** [19]) and evaluate their consistency across various retrieval sources. A key advantage of the LLM-based approach is its independence from online traffic data. While conventional metrics often rely on future engagement signals to accurately measure precision/recall, our metric operates strictly offline, providing a high-fidelity relevance signal without the need for real-time interaction logs.

3.1 Online Conversion Precision/Recall

Utilizing online traffic, we employ a look-back evaluation strategy that links observed conversions to their contributing ad requests [19]. We define relevance using a strict exact-match criterion, where a retrieved product must be identical to the converted item within a one-day post-impression attribution window. This focus assesses the system’s ability to surface the precise product driving a transaction. Accordingly, Conversion Precision (**CP**) is the proportion of retrieved products that match the conversion, while Conversion Recall (**CR**) measures the fraction of ultimately converted products successfully surfaced by the system. Metrics are limited to users with both an eligible ad request and a subsequent confirmed conversion.

3.2 Retrieval Source Baselines

A production DPA retrieval system blends several heterogeneous sources, each with distinct algorithmic and signal-coverage profiles, so a single offline metric that ranks them consistently must generalize across signal types. To make our evaluation reproducible and to expose where each source is expected to win or lose, we describe the six representative retrieval sources used throughout this study. **Matched Features (MF)** is a deterministic retargeting source: candidates are products whose canonical IDs appear in the user’s recent engagement log, optionally expanded by the seller’s product-graph siblings. Coverage is narrow but precision is high by construction. **Similar User (SU)** is a two-tower collaborative model trained on (u, p) engagement pairs with sampled softmax; at serving time the user tower’s vector is used to retrieve the top items from an ANN index over the global active catalog. **Content-Based Retrieval (CBR)** encodes each catalog item with an in-house multimodal encoder (image + title + structured attributes) and serves products that are nearest neighbors of the user’s recently engaged items in this content embedding space. **Engagement-Based Retrieval (EBR)** is a sequence-aware retrieval tower fine-tuned on impression-attributed engagement; it shares the same ANN backend as CBR but its encoder is trained on behavioral co-occurrence rather than content similarity, so it captures complementary patterns such as cross-category purchase pairs. **Popularity-Based Retrieval (PBR)** ranks the global catalog by a smoothed conversion rate over the trailing N_d , segmented by coarse demographic buckets; it has no per-user signal beyond the segment. The **Random Baseline (Rand)** samples products uniformly from a coarse category bucket aligned with the user’s last engaged category. All sources retrieve a top- $K=15$ candidate set per request; the same downstream ranker and budget policy is applied so observed differences are attributable to the retrieval stage.

3.3 Alignment Between User-Product Relevance and Conversion-Based Metrics

Table 1 reports the LLM-based offline metrics alongside conversion precision/recall (CP, CR) over a 30-day production window. NDCG@15 and MAP@15 are rank-aware retrieval metrics computed on the same LLM-judge labels, while the remaining columns follow the convention in §2.3. The two sets of rankings agree closely, indicating that the LLM judge tracks conversion-based outcomes. We also report *ImpRel*, a metric derived from impressions selected

Table 1: Retrieval source performance, reported as *log-rel* against the Random Baseline, i.e., $\log_{10}(\text{method}/\text{Rand})$ measures relative improvement (higher is better). And () indicates the ranking across all retrieval sources.

Src	CP	CR	NDCG@15	MAP@15	MicroRel	MacroRel	ImpRel
MF	+3.13 (1)	+0.91 (1)	+0.50 (1)	+0.29 (1)	+1.21 (1)	+1.02 (1)	+0.60 (1)
SU	+2.04 (2)	+0.37 (2)	+0.16 (4)	+0.19 (2)	+0.75 (2)	+0.30 (2)	+0.39 (2)
PBR	+0.45 (3)	+0.59 (3)	-0.03 (6)	+0.03 (4)	+0.12 (5)	-0.08 (6)	+0.18 (4)
CBR	+0.15 (4)	+0.30 (4)	+0.35 (2)	+0.11 (3)	+0.32 (3)	+0.07 (3)	+0.23 (3)
EBR	+0.15 (5)	+0.29 (5)	+0.28 (3)	+0.01 (5)	+0.24 (4)	+0.05 (4)	+0.12 (5)
Rand	0.00 (6)	0.00 (6)	0.00 (5)	0.00 (6)	0.00 (6)	0.00 (5)	0.00 (6)

Table 2: Offline vs. online metrics for two retrieval families (content-based, engagement-based) across four online experiments (two for each family). *log-rel* reports each version against its controls, so positive values across both rows reflect successive improvements. All four online CTR/CVR deltas are statistically significant under our internal A/B framework within 14 days of windows with 10% of DPA traffics.

Retrieval Source	Offline LLM Judge		Online A/B Metrics	
	MicroRel	MacroRel	CTR	CVR
Content-based Retrieval Model 1	+0.138	+0.122	+0.149%	+1.472%
Content-based Retrieval Model 2	+0.176	+0.169	+0.211%	+4.98%
Engagement-based Retrieval Model 1	+0.068	+0.019	+0.350%	+2.604%
Engagement-based Retrieval Model 2	+0.062	+0.095	+0.210%	+2.194%

by the downstream click prediction model from the initial retrieval pool, which captures end-to-end performance. The rankings mostly agree: CP and MicroRel produce the same ordering except for Popularity-Based Retrieval, which carries global engagement trends rather than semantic content correlation. Pearson correlations between MicroRel and CP/CR/CVR under the $k=7$ aggregation default are reported in Figure 2. Across the six retrieval sources, MicroRel achieves the highest cross-source Pearson with CP ($r=0.97$), with ImpRel ($r=0.95$) and MacroRel ($r=0.93$) close behind. Standard rank-aware retrieval metrics computed on the same LLM-judge labels are weaker proxies for the production signal: MAP@15 reaches $r=0.94$ and NDCG@15 only $r=0.55$. Intent-aware aggregation therefore tracks online conversion better than off-the-shelf rank-aware metrics.

3.4 Alignment Between User-Product Relevance and Online Metrics

A natural follow-up: *do offline gains in the relevance metric translate into online wins?* Table 2 shows directional agreement between offline relevance improvements and online CTR/CVR gains, so the offline metric is usable for guiding DPA retrieval model development before paying for an A/B test. In this experiment, we analyze two primary retrieval categories within our DPA system: content-based and engagement-based retrieval. We evaluate the LLM-based relevance scores across several iterations of these models to assess performance improvements. We then further validate them through online A/B experiments in the DPA online system. In Table 2, Model 1 and Model 2 are two successive iterations of each retrieval family; each row reports its version against its prior baseline

Table 3: Intent-set construction ablation at $k = 7$. Column definitions: see §4.1.

Intent-set Variant	Cost	Spear. vs. Full	P(CVR)	P(CR)	P(CP)
Full sequence (baseline)	100%	1.000	0.24	0.19	0.32
Random subsample (25% L_seq)	25%	0.905	0.26	0.23	0.19
Last 1-day window	11%	0.528	0.21	0.25	0.31
SID only, no event split	3.4%	0.990	0.30	0.17	0.24
SID + event split, uniform	3.4%	0.991	0.29	0.18	0.26
SID + event split, weighted	3.4%	0.989	0.50	0.29	0.35

(Model 1 vs. the preceding production model; Model 2 vs. Model 1) using $\log\text{-rel} = \log_{10}(\text{NewVersionRelevance}/\text{BaseVersionRelevance})$, so positive values consistently indicate improvement. Interestingly, engagement-based retrieval exhibits relatively weak correlations, likely because the enhanced capability reflects a blend of engagement signals and content features.

What changed across iterations. The two retrieval families were iterated along orthogonal axes; within each family, the same axes apply to both Model 1 (over the preceding production model) and Model 2 (over Model 1). Each content-based iteration enlarges the training corpus to broaden catalog coverage and refines the seed-product selection strategy used to construct (seed, candidate) training pairs, together improving content alignment for under-represented product types. Each engagement-based iteration strengthens alignment with the downstream product ranker by jointly tuning their objectives and incorporates more engagement signals directly correlated with CVR (e.g., upweighting conversion-attributed events). Because the engagement-based improvement mixes ranker-alignment with the underlying engagement signal, its offline-online delta is smaller and noisier, which explains the weaker correlation in Table 2.

4 Ablation Study

We construct a 10K dataset from retrieval logs, audited through multiple rounds of human review to reach consensus. To ensure reliable labels, we adopt a multi-agent framework in which GPT-5 and Gemini-2.5-Pro independently generate judgments that are aggregated via confidence scores; human annotators resolve disagreements until convergence, and the final judgments are treated as ground truth. Results of the main ablation are shown in Table 4. In addition, we ablate two design choices that are central to the framework but whose empirical justification would have interrupted the methodology in §2.1 and §2.2.1: the SID-based intent-set construction (§4.1), and the two-phase agentic-then-human guideline-generation procedure (§4.2).

4.1 Intent-set construction

We ablate the intent-set decomposition along three axes – (i) random vs. recency-based vs. SID-based selection, (ii) presence or absence of event-type stratification, and (iii) uniform vs. size-and-event-weighted aggregation – to isolate the contribution of each design choice rather than of subsampling alone. To put all variants on the same label matrix, we exhaustively judge every event–candidate pair in the 30-day window with the teacher LLM, then for each variant compute its offline MicroRel and report the Pearson correlation against per-(date, source) live CVR, CR, and CP. In Table 3, *Cost* is pairwise calls as a percentage of the raw 30-day history (several

Table 4: Impact of Guidelines and Student Models

Role	Model Config	Precision	Recall	F1
Teacher	Gemini-2.5-Pro (no guidel.)	0.824	0.825	0.816
	Gemini-2.5-Pro (with guidel.)	0.861	0.866	0.860
Student	SigLIP-2-mm [16]	0.780	0.751	0.746
	Gemini-2.5-Flash (with guidel.)	0.840	0.843	0.832
	Gemini-2.5-Flash-ft	0.869	0.873	0.869
	Gemini-2.5-Flash-lite-ft	0.845	0.849	0.846

hundred events per user after product-id deduplication), placing all variants on a common denominator; *Spear. vs. Full* is the per-user MicroRel rank correlation against the full sequence; and $P(\cdot)$ is the Pearson between offline MicroRel and the named online metric. Random baselines average over multiple seeds.

Two findings stand out. First, recency truncation halves intent-set fidelity (Last 1-day Spearman 0.53 vs. Full’s 1.00) and yet barely moves online Pearson – a Last 1-day intent set tracks CP/CR/CVR within ± 0.03 –0.06 of the Full sequence. The signal that drives online business metrics is concentrated in a small, recent slice of engagement, even though the intent-set structure looks very different. This is what makes aggressive compression viable in the first place. Second, SID grouping with event-type stratification keeps that compression while also restoring the intent-set fidelity: it matches the full sequence at 0.99 Spearman at only 3.4% of its pairwise cost – a $\sim 30\times$ reduction in LLM calls – and the explicit $w_g = \log(1+|g|) \cdot \beta(e_g)$ weighting lifts the variant to $P(\text{CVR}) = 0.50$, $P(\text{CR}) = 0.29$, $P(\text{CP}) = 0.35$, beating Full and every other variant on all three online metrics. The margin over the Random subsample baseline (which itself costs $7\times$ more) is $+0.24$ on CVR and $+0.16$ on CP. We attribute this to a regularization effect: by deduplicating near-identical events and explicitly downweighting low-intent views, the metric stops rewarding retrieval algorithms that re-surface repeated views of the same item – a failure mode the full-sequence MicroRel cannot distinguish from genuine relevance.

4.2 Impact of Guidelines

Using Gemini-2.5-Pro as the backbone, a judge given only brief relevance definitions (without the guideline) is 4% worse on both precision and recall than the guideline-prompted version (Table 4) and even underperforms the smaller student model with guideline, showing that the guideline carries most of the lift. The gain concentrates on rule classes that the loose prompt systematically misjudges, e.g., demographic mismatches and brand-vs.-category confusions, which the agentic discovery loop in §2.2.1 surfaces.

4.3 Choice of Student Judge

We adopt Gemini-2.5-Flash as the student backbone. Instruction fine-tuning improves precision by 2.9% and recall by 3.0%, outperforming detailed prompting alone and fine-tuned Gemini-2.5-Flash-Lite. The fine-tuned Flash student in fact matches the Pro teacher on F_1 (0.869 vs. 0.860) while running at a fraction of the cost (Table 5), so we deploy it as the production judge $\mathcal{J}$. It surpasses SigLIP2-MM—which uses $10\times$ more training data and a two-layer classifier over fused intent–product embeddings—by 9% in precision and 12%

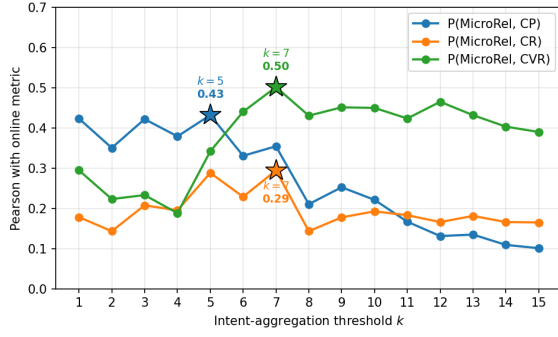


Figure 2: MicroRel’s Pearson correlation with live CP, CR, and CVR as a function of the intent-aggregation threshold k . Stars mark per-metric peaks; all three peaks fall in $k \in [5, 7]$. At the $k=7$ default, every CI is strictly positive.

in recall. The lightweight classifier struggles with complex relevance patterns learning from limited training data, highlighting the advantage of LLM-based judges with few-shot learning.

4.4 Intent-Aggregation Threshold k

We compare the production default $k=7$ against the at-least-one permissive baseline ($k=1$), which equates a single weak intent match with broad intent coverage. We sweep $k \in [1, 15]$ on the same exhaustive event–candidate judgment matrix used in §4.1 so the only varying quantity is the aggregation rule. Figure 2 reports the Pearson correlation between the resulting MicroRel and CP, CR, and CVR measured per-(date, source) over a 30-day production window across 2 retrieval sources. All three correlations reach their maxima within $k \in [5, 7]$: P(CP) peaks at 0.43 ($k=5$), and P(CR) and P(CVR) both peak at $k=7$. Beyond $k \geq 8$, P(CP) and P(CR) decay as the candidate pass-rate drops below $\sim 5\%$, while CVR retains most of its signal through $k=13$. At the $k=7$, all three 95% bootstrap CIs are strictly positive: P(CP) $[+0.04, +0.63]$, P(CR) $[+0.02, +0.55]$, P(CVR) $[+0.29, +0.68]$. The intuition is that high- k aggregation filters out single weak intent matches that are merely category-coincidental, requiring multi-intent support before a candidate counts as relevant, which aligns better with multi-touch purchase behavior.

5 System Efficiency

We benchmark four judge configurations on Vertex AI at 16-thread concurrency over 100 production retrieval pairs (Table 5). The fine-tuned Gemini-2.5-Flash we deploy reaches 751 ms p50, 1,117 ms p95, and 18 QPS – roughly $9\times$ faster than the teacher on p50 latency and on throughput. The fine-tuned Gemini-2.5-Flash-Lite variant offers a marginal further gain at 647 ms p50 and 20 QPS. Most of the latency improvement therefore comes from fine-tuning rather than from picking a smaller backbone: prompt-only Flash still inherits the teacher’s long-tail p95 (8,807 ms), whereas the fine-tuned student compresses both p50 and p95 by an order of magnitude.

6 Related Work

Offline evaluation for recommendation and ads retrieval. Classical offline metrics for recommendation retrieval – Recall@ K , NDCG@ K over future-engagement labels, and conversion-attributed

Table 5: System efficiency of teacher and student judges. Latency is per-pair; QPS is aggregate throughput at the listed concurrency.

Judge	p50 (ms)	p95 (ms)	QPS
Gemini-2.5-Pro (teacher, with guidel.)	6,963	10,840	2
Gemini-2.5-Flash (prompt, with guidel.)	2,189	8,807	4
Gemini-2.5-Flash-ft (ours)	751	1,117	18
Gemini-2.5-Flash-Lite-ft	647	1,342	20

precision/recall [19] – have well-known weaknesses in ads: severe selection bias from the exposure policy, heavy tail of low-engagement requests, and a lookback window that conflates retrieval quality with downstream ranker and budget effects. The offline-online evaluation gap in e-commerce recommendation has been documented at multiple scales [8]. Counterfactual estimators (IPS, doubly robust) reduce some of this bias [5] but assume access to logged propensities that are rarely available at the retrieval stage of a multi-tower ad system. Our work is complementary: rather than debiasing engagement labels, we replace them with a content-grounded relevance label produced by an LLM judge, and report alignment with engagement-based offline metrics.

LLM judges for relevance. LLM-as-judge has moved from open-domain summarization [14] to e-commerce and recommender relevance labeling: LREF [15] fine-tunes an LLM as a query-product relevance scorer with chain-of-thought, BloomIntent [2] performs intent-conditioned evaluation of SERP satisfaction, Wang et al. [18] progressively train an LLM-based semantic-ads retrieval encoder at Walmart, Penha et al. [7] examine LLM-judge ranking agreement with humans in Cranfield-style movie and podcast recommendation, and large-scale studies show LLM labels broadly track human judgment in IR settings [10, 17]. Two gaps remain. *First*, these works rely on either an explicit query (search) or a Cranfield-style topic per request (movie/podcast collections); DPA retrieval has neither – intent must be inferred from heterogeneous engagement signals, so we propose SID-grouped, event-stratified intent decomposition as the compression rule over the engagement history. *Second*, prior LLM-judge studies validate against *human assessors* at the pair or ranking level; we instead validate against *online business KPIs* (CTR, CVR, conversion P/R) on the launch-decision critical path.

7 Conclusion and Limitation

We present a query-free user–product relevance metric for offline evaluation of DPA retrieval. A user’s engagement history is compressed into a small SID-grouped, event-stratified intent set, and an LLM judge scores each retrieved candidate against this set; an agentically-discovered guideline anchors a Gemini-2.5-Pro teacher, distilled into a Gemini-2.5-Flash student that matches the teacher at $\sim 9\times$ lower latency. On 30 days of production DPA traffic, the metric ranks six retrieval sources consistently with conversion precision/recall and is directionally aligned with online CTR/CVR across four A/B tests; it serves as the offline gate for DPA model evaluation before online deployment.

8 GenAI Usage Disclosure

In accordance with the ACM Policy on the Use of Generative AI Tools and Technologies, we disclose that GenAI assistants (including ChatGPT and Claude) were used during the preparation of the manuscript for language editing and writing assistance, including improving grammar, clarity, readability, and consistency of terminology. We also used AI-assisted tools to help draft preliminary versions of certain figures and diagrams, which we subsequently reviewed, revised, and finalized. All research ideas, methodology, system design, experiments, analyses, results, and conclusions were developed and validated by the authors. All cited references were manually identified by the authors and verified against their original sources with the help of AI tools. Because this paper studies LLM-based relevance judge, the use of LLMs as components of the studied systems and experimental methodology is described separately in the relevant sections of the paper.

References

- [1] Christine Bauer, Eva Zangerle, and Alan Said. 2024. Exploring the landscape of recommender systems evaluation: Practices and perspectives. *ACM Transactions on Recommender Systems* 2, 1 (2024), 1–31.
- [2] Yoonseo Choi, Eunhye Kim, Hyunwoo Kim, Donghyun Park, Honggu Lee, Jin Young Kim, and Juho Kim. 2025. BloomIntent: Automating Search Evaluation with LLM-Generated Fine-Grained User Intents. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–34.
- [3] Jiaxin Deng, Shiyao Wang, Kuo Cai, Lejian Ren, Qigen Hu, Weifeng Ding, Qiang Luo, and Guorui Zhou. 2025. Onerec: Unifying retrieve and rank with generative recommender and iterative preference alignment. *arXiv preprint arXiv:2502.18965* (2025).
- [4] Byung Eun Jeon, Ryan Bae, and Xiao Bai. 2025. Leveraging Large Language Models for Complementary Product Ads Recommendation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. 4837–4841.
- [5] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2017. Accurately interpreting clickthrough data as implicit feedback. In *Acm Sigir Forum*, Vol. 51. Acm New York, NY, USA, 4–11.
- [6] Smitha Milli, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao, and Anca D Dragan. 2025. Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS nexus* 4, 3 (2025), pgaf062.
- [7] Gustavo Penha, Aleksandr V Petrov, Claudia Hauff, Enrico Palumbo, Ali Vardasbi, Edoardo D’Amico, Francesco Fabbri, Alice Wang, Praveen Chandar, Henrik Lindstrom, Hugues Bouchard, and Mounia Lalmas. 2025. Do LLM-judges Align with Human Relevance in Cranfield-style Recommender Evaluation? *arXiv preprint arXiv:2511.23312* (2025).
- [8] Ladislav Peska and Peter Vojtas. 2020. Off-line vs. On-line Evaluation of Recommender Systems in Small E-commerce. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*. 291–300.
- [9] Pearl Pu, Li Chen, and Rong Hu. 2012. Evaluating recommender systems from the user’s perspective: survey of the state of the art. *User Modeling and User-Adapted Interaction* 22, 4 (2012), 317–355.
- [10] Hossein A Rahmani, Clemencia Siro, Mohammad Aliannejadi, Nick Craswell, Charles LA Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz. 2025. Judging the judges: A collection of llm-generated relevance judgements. *arXiv preprint arXiv:2502.13908* (2025).
- [11] Chandan K Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. 2022. Shopping queries dataset: A large-scale ESCI benchmark for improving product search. *arXiv preprint arXiv:2206.06588* (2022).
- [12] Guy Shani and Asela Gunawardana. 2015. Evaluating recommender systems. *Recommender Systems Handbook* (2015), 257–297.
- [13] Anima Singh, Trung Vu, Nikhil Mehta, Raghunandan Keshavan, Maheswaran Sathiamoorthy, Yilin Zheng, Lichan Hong, Lukasz Heldt, Li Wei, Devansh Tandon, et al. 2024. Better generalization with semantic ids: A case study in ranking for recommendations. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 1039–1044.
- [14] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. *arXiv preprint arXiv:2304.09542* (2023).
- [15] Tian Tang, Zhixing Tian, Zhenyu Zhu, Chenyang Wang, Haiqing Hu, Guoyu Tang, Lin Liu, and Sulong Xu. 2025. LREF: A Novel LLM-based Relevance Framework for E-commerce Search. In *Companion Proceedings of the ACM on Web Conference 2025*. 468–475.
- [16] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786* (2025).
- [17] Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Daniel Campos, Nick Craswell, Ian Soboroff, Hoa Trang Dang, and Jimmy Lin. 2024. A large-scale study of relevance assessments with large language models: An initial look. *arXiv preprint arXiv:2411.08275* (2024).
- [18] Zhaodong Wang, Weizhi Du, Md Omar Faruk Rokon, Pooshpendu Adhikary, Yanbing Xue, Jiaxuan Xu, Jianghong Zhou, Kuang-chih Lee, and Musen Wen. 2025. Semantic Ads Retrieval at Walmart eCommerce with Language Models Progressively Trained on Multiple Knowledge Domains. *arXiv preprint arXiv:2502.09089* (2025).
- [19] Jiang Zhang, Sumit Kumar, Wei Chang, Yubo Wang, Feng Zhang, Weize Mao, Hanchao Yu, Aashu Singh, Min Li, and Qifan Wang. 2025. Optimizing Recall or Relevance? A Multi-Task Multi-Head Approach for Item-to-Item Retrieval in Recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 5194–5204.