Scaling Dense Retrieval with LLM-Annotated Training Data: Structured Mining and Progressive Curriculum for E-Commerce Sponsored Search

Md Omar Faruk Rokon[1,*], Shasvat Desai[1,*], Jhalak Nilesh Acharya[1,*], Isha Shah[1], Kumar Priyam[1], Brahanyaa Somasundaram[1], Vamsee Tangirala[1], Minuteresa Thomas[1], Vivek Arora[1], Vijay Manchi[1], Hong Yao[1] and Kuang-chih Lee[1]

[1] *Walmart Global Tech, Sunnyvale, CA, USA*

Abstract

How can we generate high-quality training data for dense retrieval models at production scale, without relying on click signals or manual annotation? This question is critical for e-commerce sponsored search, where click-based training suffers from position bias and tail-query sparsity, and manual labeling at the scale of hundreds of millions of query-item pairs is economically infeasible. Our work is driven by the following insight: heterogeneous retrieval systems disagree on most items they retrieve, and this disagreement creates a natural source of structured training signal—easy positives where all systems agree, hard positives that only lexical systems find, and hard negatives that fool exactly one system. As our key novelty, we combine three ideas into an end-to-end pipeline: (a) multi-channel retrieval mining with rank metadata from three production systems, (b) graded-relevance annotation by a calibrated three-model cascade (184M cross-encoder → LoRA-adapted 2B LLM → LoRA-adapted 8B LLM) that reaches 89.1% agreement with trained human annotators, and (c) three-stage progressive curriculum training (BCE → MNR → Triplet) that organizes 240M+ training examples across five difficulty levels. We deploy the trained two-tower BERT model on Walmart's sponsored search and evaluate it against 30K queries labeled by trained third-party human annotators. First, we show that the system achieves +5.1% NDCG@10 (0.878 → 0.923) over the click-trained production baseline, with the largest gain on tail queries (+6.8%). Second, we show that embarrassing retrievals (rating 0) drop from 8.7% to 3.5%. Third, a two-week online A/B test with tens of millions of ad requests per arm confirms +2.80% ad spend, +1.4% CTR, +2.8% eCPM, and +2.9% click conversion rate. Overall, our work provides a practical and scalable blueprint for replacing click-based training with structured LLM-annotated supervision in production retrieval systems.

Keywords

Dense Retrieval, LLM Annotation, Distant Supervision, Curriculum Learning, Training Data Mining, E-Commerce Search

1. Introduction

How can we train dense retrieval models for e-commerce sponsored search at production scale, when click signals are biased and manual annotation is too expensive? This question is important because retrieval is the first stage in a sponsored search pipeline: it determines which ads a user sees, and everything downstream—ranking, bidding, revenue—depends on its quality [1, 37, 38]. On platforms like Walmart.com, even small improvements in retrieval relevance translate to substantial business outcomes.

Consider a concrete example. When a user searches for "iPhone charger," the retrieval model must surface relevant sponsored items such as Lightning cables and USB-C adapters, while avoiding embarrassing results like phone cases or screen protectors. A click-trained model learns which items users clicked on for this query in the past—but clicks are noisy. Users click top-ranked results regardless

ECOM'26: SIGIR Workshop on eCommerce, Jul 24, 2026, Melbourne, Australia

*Corresponding author.

mdomarfaruk.rokon@walmart.com (M. O. F. Rokon); shasvat.desai@walmart.com (S. Desai); jhalak.acharya@walmart.com (J. N. Acharya); isha.shah@walmart.com (I. Shah); kumar.priyam@walmart.com (K. Priyam); brahanyaa.somasundaram@walmart.com (B. Somasundaram); vamsee.tangirala@walmart.com (V. Tangirala); minuteresa.thomas@walmart.com (M. Thomas); vivek.arora@walmart.com (V. Arora); vijay.manchi@walmart.com (V. Manchi); hong.yao0@walmart.com (H. Yao); kuang-chih.lee@walmart.com (K. Lee)

of relevance (position bias [37, 33]), and for tail queries like "MagSafe charger for iPhone 15 Pro Max," there may be no click history at all (the cold-start problem). Manual annotation can produce clean labels, but at the scale we need—240M+ query-item pairs across 4M queries—it would cost tens of millions of dollars per refresh cycle, which is infeasible for weekly model retraining.

Our work is driven by the following insight: the three retrieval systems already running in production—a dictionary-based lexical system, a BM25 system, and the current ANN embedding model—disagree on most of the items they retrieve. At the top 500, their pairwise overlap is only 13–15%. This disagreement is not noise; it is structured signal. Items that all three systems retrieve and that an LLM judges relevant are high-confidence easy positives. Items that lexical systems find but the ANN misses are hard positives—exactly the failures we want the next model to fix. Items retrieved by only one system and judged irrelevant are hard negatives that confuse real production systems. By combining multi-channel disagreement with a calibrated LLM annotation cascade, we can mine millions of structured training examples at five distinct difficulty levels, without any click signals or manual labels.

As our key novelty, we combine three ideas into a single end-to-end pipeline: (a) structured data mining from multi-channel retrieval disagreement, using rank position, channel agreement, and catalog-level token similarity to create five difficulty levels of training examples; (b) a calibrated annotation cascade of three fine-tuned classifiers (184M cross-encoder $\rightarrow$ LoRA-adapted 2B LLM $\rightarrow$ LoRA-adapted 8B LLM) with per-class isotonic calibration, reaching 89.1% agreement with trained human annotators—extending our prior work on single-model binary annotation [31] to a multi-model cascade on 5-class graded relevance; and (c) progressive curriculum training across three stages of increasing difficulty (BCE $\rightarrow$ MNR $\rightarrow$ Triplet), where each stage uses a loss function matched to the discrimination granularity required by its sample types. The pipeline generates 240M+ training examples across 4M queries and enables fully automated weekly model retraining. The approach is a form of distant supervision [21]: the cascade is used once, offline, to produce labels; the student is then trained with standard supervised objectives. We intentionally do not perform classical knowledge distillation [12]—no soft-label transfer, no teacher gradient flow, no co-training—because decoupling the annotator from the student means annotations can be cached, reused, and the two components upgraded independently.

We deploy the resulting model (Embedding Model V3) on Walmart's sponsored search and evaluate it against 30K queries labeled by trained third-party human annotators—independent of the cascade used for training. First, we show that the system achieves +5.1% NDCG@10 (0.878 $\rightarrow$ 0.923) over the click-trained production baseline, with embarrassing results reduced from 8.7% to 3.5%. Second, we show that improvements are consistent across all query segments, with the largest gain on tail queries (+6.8% NDCG@10), consistent with addressing the cold-start weakness of click-trained models. Third, a two-week online A/B test with tens of millions of ad requests per arm suggests that these offline gains transfer to online business metrics: +2.80% ad spend, +1.4% CTR, +2.8% eCPM, and +2.9% click conversion rate.

Our contributions are: (a) a structured data mining pipeline that converts multi-channel retrieval disagreement into five difficulty levels of training examples—the key upstream innovation that lets a standard two-tower BERT achieve production-quality retrieval without architectural changes; (b) a calibrated three-model annotation cascade with per-class isotonic calibration that halves compute relative to running all models on every pair, while maintaining 89.1% agreement with human annotators on a 5-class graded-relevance task; (c) a three-stage progressive curriculum (BCE $\rightarrow$ MNR $\rightarrow$ Triplet) that yields +9.5% NDCG@10 over single-stage training on the same data; and (d) production deployment and online A/B validation at Walmart scale, providing evidence that offline relevance gains transfer to measurable business outcomes.

The remainder of this paper is organized as follows. Section 2 reviews related work on neural retrieval, training data generation, and LLM-based annotation. Section 3 defines the problem and the training data challenge. Section 4 describes our four-stage pipeline in detail. Section 5 presents the experimental setup, and Section 6 reports results. Section 7 discusses findings and limitations, and Section 8 concludes.

2. Related Work

Our work sits at the intersection of four research areas: neural retrieval architectures, training data generation, LLM-based annotation, and curriculum learning.

Neural retrieval for e-commerce. The two-tower architecture was introduced by Huang et al. [15] as DSSM and later adopted for YouTube recommendations [5]. It has since become the dominant paradigm for large-scale retrieval [17]. In e-commerce, Nigam et al. [23] built semantic product search at Amazon, Huang et al. [14] deployed embedding-based retrieval at Facebook, and Fan et al. [10] proposed MOBIUS for Baidu's sponsored search. At Walmart, Magnani et al. [20] introduced semantic retrieval for product search, Wang et al. [38] proposed progressive fusion training for ads retrieval, and Desai et al. [7] created a unified supervision framework using relevance and engagement signals. Our work does not propose a new model architecture. Instead, we address the upstream problem of training data quality, which is orthogonal to these architectural advances.

General-purpose dense retrievers. Recent work has produced strong general-purpose embedding models such as E5 [36], BGE [39], and GTE [18], which achieve impressive zero-shot performance on the BEIR benchmark [34]. Learned sparse approaches like SPLADE [11] offer a complementary efficiency profile. These models are strong off-the-shelf baselines, but our production setting imposes constraints they do not address out of the box: sponsored-ad catalog matching, strict serving latency requirements with ANN indexing, and weekly refresh on a shifting query distribution. This is why we train a domain-adapted student rather than serving a frozen general retriever.

Distillation and distant supervision for retrieval. Knowledge distillation [12] transfers a teacher's soft label distribution to a smaller student. In retrieval, Hofstätter et al. [13] proposed cross-architecture distillation for efficient ranking, Qu et al. [24] introduced RocketQA with cross-encoder supervision, and Reimers and Gurevych [28] distilled monolingual embeddings into multilingual models. All of these methods require the teacher during student training. Our approach is different: it is better described as distant supervision [21]. The teacher (LLM cascade) is used once, offline, to produce hard labels, and the student is trained with standard supervised objectives. This is conceptually similar to weakly-supervised pre-training [36, 39] and LLM-generated training data [3, 6], but differs in its source—we mine real production query-item pairs rather than generating synthetic queries—and in its scale and production integration.

Training data generation for retrieval. Generating high-quality training data at scale is a long-standing challenge. Bonifacio et al. [3] used LLMs for synthetic query generation, and Dai et al. [6] proposed Promptagator for few-shot dense retrieval via LLM-generated queries. Reddy et al. [26] released the Shopping Queries Dataset (ESCI) with multi-class relevance labels for e-commerce retrieval evaluation. Unlike these approaches, we do not generate synthetic data. Instead, we mine real production query-item pairs from multiple retrieval channels and annotate them with calibrated LLM labels, producing training data that directly reflects the production query distribution.

LLMs as relevance judges. Recent work shows that LLMs can serve as effective relevance judges. Thomas et al. [35] showed that LLMs accurately predict searcher preferences, Faggioli et al. [9] surveyed LLMs for IR relevance judgment, and Zheng et al. [42] characterized the biases and strengths of LLM-as-judge protocols. In e-commerce, Rokon et al. [31] showed that a single LoRA-adapted LLM reaches 89.43% accuracy on binary query-ad relevance classification. Our work extends that line in three ways: (1) we move from binary to 5-class graded relevance, which is harder and more useful for ranking; (2) we introduce a three-model cascade with per-class isotonic calibration [4] that halves compute at equivalent accuracy; and (3) we integrate this cascade into a larger pipeline with multi-channel mining, curriculum training, and production deployment.

Hard negative mining and curriculum learning. The quality of negative examples is critical for contrastive training [40, 41, 30]. ANCE [40] proposed asynchronous index-based hard negative sampling, and Zhan et al. [41] studied optimal hard negative selection strategies. Curriculum learning [2] advocates organizing training data from easy to hard. Our approach combines both: we mine multiple difficulty levels of negatives from multi-channel disagreement patterns and train across three stages of increasing difficulty, each with a distinct loss function matched to the sample type.

3. Problem Formulation

3.1. Task Definition

We focus on the retrieval stage of sponsored search. Given a user search query $q \in \mathcal{Q}$, the goal is to identify a set of relevant sponsored items $\mathcal{I}_q \subset \mathcal{C}$ from the product catalog $\mathcal{C}$, such that items in $\mathcal{I}_q$ are relevant to q and have active advertiser bids. We measure relevance on a five-point graded scale [16]:

$$\text{rel} : \mathcal{Q} \times \mathcal{C} \to \{0, 1, 2, 3, 4\} \tag{1}$$

where 0 = Embarrassing, 1 = Bad, 2 = Okay, 3 = Good, and 4 = Excellent. Items with $\text{rel}(q, i) \geq 3$ are considered relevant for retrieval purposes. The retrieval model must serve results within strict latency constraints suitable for real-time search.

3.2. The Training Data Challenge

The core challenge is generating training data at the scale and quality needed for weekly model retraining. There are three bottlenecks.

Clicks are biased. The probability of a click depends on display position, not just relevance: $P(\text{click} \mid q, i, \text{pos}) \neq P(\text{rel} \mid q, i)$. Users click top-ranked results regardless of quality [37, 33], so models trained on clicks learn to replicate the existing system's ranking rather than true relevance.

Clicks are sparse. For tail queries, clicks approach zero: $\mathbb{E}[\text{clicks}(q, i)] \to 0$ for $q \in \mathcal{Q}_{\text{tail}}$. New items and infrequent queries—which make up a significant fraction of query volume—receive no training signal at all.

Manual annotation does not scale. Human annotation produces clean labels but costs \$0.10–0.50 per pair. At the scale we need (240M+ examples across 4M queries), this would cost \$24M–\$120M per refresh cycle—infeasible for weekly retraining.

4. Method

Our pipeline has four stages (Figure 1): (1) collect retrieval results with rank information from three production channels, (2) annotate all query-item pairs with a calibrated LLM cascade, (3) classify pairs into five difficulty levels using channel agreement and rank thresholds, and (4) train the student model through a three-stage curriculum. We describe each stage below.

4.1. Retrieval Infrastructure

The data collection stage leverages three heterogeneous retrieval channels operating in parallel on Walmart's sponsored search platform, each providing a distinct view of query-item relevance:

- **Dictionary-based text retrieval (R_D)**: A BM25-variant retrieval system on the advertising catalog whose item index is augmented with historically relevant queries, enabling exact-match retrieval for known query-item associations and falling back to lexical BM25 scoring otherwise.
- **BM25-based retrieval (R_B)**: Standard BM25 [29] retrieval on the advertising catalog using item title, description, brand, and product type as index fields, without query augmentation.
- **ANN retrieval (R_A, Embedding Model V2)**: The production neural retrieval model using a two-tower BERT [8] architecture trained with click-based labels and approximate nearest neighbor search [38].

For each query q, we collect the top-K=500 retrieved items from each channel, along with each item's rank position $r_s(q, i)$ within channel s. These channels exhibit low pairwise overlap—dictionary-ANN overlap is 13.4% and dictionary-BM25 overlap is 15.05% at K=500—meaning the majority of retrieved items exist in disagreement regions. This low overlap is a feature: it provides abundant structured signal for training data extraction (Section 4.3).

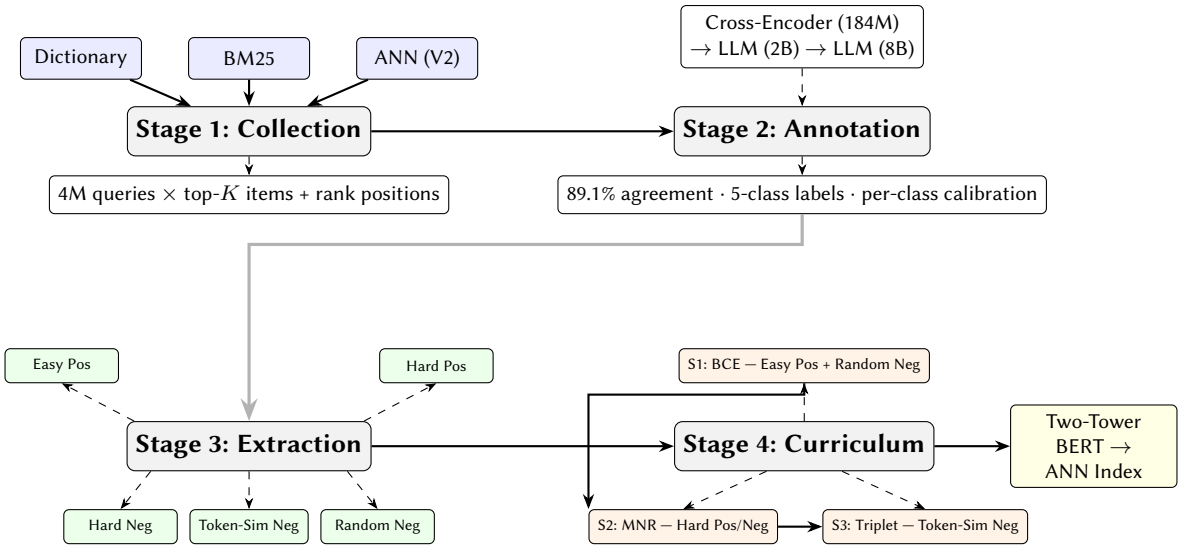

Figure 1: Overview of the training-data pipeline. **Top row:** Stage 1 collects retrieval results with rank positions from three heterogeneous channels; Stage 2 annotates all pairs via a calibrated model cascade (89.1% agreement with human annotators). **Bottom row:** Stage 3 classifies pairs into five difficulty levels using channel agreement, rank thresholds, and catalog mining; Stage 4 trains a two-tower BERT model through three curriculum stages of increasing difficulty (BCE → MNR → Triplet), with each stage matched to specific sample types. Arrows between training stages indicate checkpoint transfer.

4.2. Calibrated Cascade Annotation

The key idea of our annotation approach is simple: use a cascade of progressively larger fine-tuned classifiers to label query-item pairs with 5-class graded relevance, and calibrate each model so that confident predictions are accepted early while ambiguous pairs are deferred to larger models. The cascade produces hard (argmax) labels, not soft probabilities. This is distant supervision [21], not classical knowledge distillation [12]: the cascade is used once, offline, and the student is trained with standard supervised objectives. We make this choice deliberately because it decouples the annotator from the student: (1) annotations are computed in batch and cached; (2) the cascade adds no serving latency; and (3) cascade and student can be upgraded independently. The cost is that we discard the cascade's uncertainty on borderline pairs; Section 7 discusses this trade-off.

Cascade architecture. The annotation engine (f_T) is a calibrated model cascade that routes query-item pairs through three progressively larger domain-specific fine-tuned classifiers [4]: (1) a cross-encoder (184M parameters), (2) a LoRA-adapted 2B-parameter LLM, and (3) a LoRA-adapted 8B-parameter LLM. Each model produces a five-class relevance prediction with a calibrated confidence score. If the confidence exceeds a per-model threshold, the prediction is accepted; otherwise, the pair is deferred to the next, larger model. Pairs that remain unresolved after all three models are decided by majority vote with a max-prediction tiebreaker.

Per-class isotonic calibration. A critical design choice is calibrating each predicted class separately via isotonic regression, rather than applying a single global calibration. This captures heterogeneous confidence distributions across classes—models tend to be overconfident on extreme classes (0 and 4) and underconfident on middle classes (1–3). Per-class calibration achieves the best cascade routing accuracy compared to temperature scaling, Platt scaling, and global isotonic methods.

Annotation quality. We audit the cascade against trained human annotators on a held-out set of 100K query-item pairs disjoint from all training data. The cascade reaches 89.1% agreement on the 5-class task, with the Cross-Encoder handling 74.5% of pairs at 91.2% agreement and larger LLMs reserved for ambiguous cases—reducing compute by approximately 50% relative to running all three models on every pair. For context we also report zero-shot agreement of off-the-shelf GPT-4 (68.2%) and Claude-3 (70.1%) on the same audit set under a matched prompt template. We emphasize that this is a *fine-tuned domain cascade vs. zero-shot general-purpose LLM* comparison, not a like-for-like evaluation of LLM capability;

the relevant takeaway is that off-the-shelf use would inject substantial label noise into the pipeline, motivating our domain-adapted cascade. At the scale of 240M+ examples, the residual ∼11% label noise in cascade labels is tolerable: prior work on noise-robust deep learning [32, 22] shows that deep classifiers converge to near-clean performance when supervised with enough loosely-correlated labels, and our multi-channel agreement filter (Section 4.3) further removes the most likely false negatives.

4.3. Structured Training Data Extraction

The key idea of our data mining approach is to use channel agreement and disagreement as a signal for training difficulty. Given the annotated query-item pairs from the cascade, we classify each pair into one of five difficulty levels based on three signals: which channels retrieved the item, where it ranked, and what the cascade labeled it. The pipeline operates over 4 million queries spanning head, torso, and tail segments. Let $\mathcal{R}_D(q)$, $\mathcal{R}_B(q)$, and $\mathcal{R}_A(q)$ denote the retrieval sets from dictionary-based, BM25-based, and ANN retrieval, respectively.

Easy Positives ($\mathcal{D}_{\text{EP}}$): Items retrieved by all three channels and confirmed relevant by the cascade—high-confidence positives where all systems agree:

$$\mathcal{D}_{\text{EP}} = \{(q, i) : i \in \mathcal{R}_D \cap \mathcal{R}_B \cap \mathcal{R}_A,\ \text{rel}(q, i) \geq 3\} \tag{2}$$

Hard Positives ($\mathcal{D}_{\text{HP}}$): Items missed by the ANN retrieval but ranked highly by at least one lexical channel, and confirmed relevant—these are items the current dense model fails to surface. Formally, $(q, i) \in \mathcal{D}_{\text{HP}}$ if $i \in (\mathcal{R}_D(q) \cup \mathcal{R}_B(q)) \setminus \mathcal{R}_A(q)$, there exists $s \in \{D, B\}$ with $r_s(q, i) \leq 50$, and $\text{rel}(q, i) \geq 3$.

Hard Negatives ($\mathcal{D}_{\text{HN}}$): Items ranked within the top 100 by exactly one channel but judged irrelevant. We apply a *multi-channel agreement filter*: negatives retrieved by more than one channel are removed, since cross-channel agreement suggests the item may be borderline relevant despite the cascade label. Formally, $(q, i) \in \mathcal{D}_{\text{HN}}$ if there exists exactly one channel $s \in \{D, B, A\}$ such that $i \in \mathcal{R}_s(q)$ and $r_s(q, i) \leq 100$, and $\text{rel}(q, i) \leq 2$.

Semi-Hard (Token-Similar) Negatives ($\mathcal{D}_{\text{SHN}}$): Items from the catalog with moderate token overlap (TF-IDF similarity $\in [0.2, 0.6]$) but *not retrieved by any channel*, additionally filtered by the cascade to remove false negatives. These lexically plausible but semantically irrelevant examples force the model to learn beyond keyword matching:

$$\mathcal{D}_{\text{SHN}} = \{(q, i) : \text{TF-IDF}(q, i) \in [0.2, 0.6],\ i \notin \mathcal{R}_D \cup \mathcal{R}_B \cup \mathcal{R}_A,\ \text{rel}(q, i) \leq 2\} \tag{3}$$

Easy (Random) Negatives ($\mathcal{D}_{\text{EN}}$): Randomly sampled catalog items with minimal lexical overlap, providing foundational negative signal and broad coverage.

Table 1 summarizes the composition of the resulting training dataset. After classification, we apply quality controls: (1) removal of queries where no retrieved item receives rel ≥ 3 (these queries lack sufficient positive support for contrastive training; approximately 8% of queries fall into this category, which includes both legitimate no-relevant-ad queries and potential systematic annotation failures), (2) deduplication of near-identical items within the same query-rating group, (3) rank-based filtering (top-50 for positives, top-100 for hard negatives), and (4) per-query sampling limits of 50 positives and 50 negatives. The theoretical maximum of ∼400M examples (4M × 100) is reduced to 240M+ by these filters: tail queries often yield fewer than 50 positives from channel agreement, rank constraints exclude lower-ranked items, deduplication removes near-duplicate product listings, and all-negative queries are discarded entirely. The resulting pipeline produces 240M+ examples (158M+ positives, 80M+ negatives).

4.4. Progressive Curriculum Training

The key idea of our training strategy is to present examples in order of increasing difficulty, so the model learns basic relevance patterns before tackling hard cases. Training on all 240M+ examples at once with a single loss function underperforms a staged approach [2]: hard negatives and token-similar

Table 1
Training data composition. Five sample types are extracted across 4M queries, producing 240M+ examples.

Sample Type	Source	Count	Curriculum
Easy Positives	3-channel agreement	158M+	Stage 1
Hard Positives	Lexical \ ANN		Stage 2–3
Hard Negatives	Single-channel, top-100	80M+	Stage 2
Token-Sim. Negatives	Catalog TF-IDF mining		Stage 3
Random Negatives	Catalog sampling		Stage 1

negatives are only useful after the model can distinguish obviously relevant from obviously irrelevant items. We train the two-tower BERT model in three stages, where each stage uses a different loss function matched to the difficulty of its sample types (Table 1). Each stage initializes from the best checkpoint of the previous stage.

Stage 1: Binary Cross-Entropy (Foundation). The model learns basic relevance discrimination using only the highest-confidence positives ($\mathcal{D}_{\text{EP}}$, rating = 4, i.e. "Excellent") and random negatives ($\mathcal{D}_{\text{EN}}$). We deliberately exclude rating-3 ("Good") items from this stage: including borderline positives alongside trivially-distinguishable random negatives would introduce label ambiguity before the model has learned any relevance signal, harming early convergence. BCE is appropriate here because the positive-negative distinction is unambiguous, requiring only pointwise scoring:

$$\mathcal{L}_{\text{BCE}} = -\big[y \cdot \log(\sigma(\tau \cdot s(q,i))) + (1-y) \cdot \log(1-\sigma(\tau \cdot s(q,i)))\big] \tag{4}$$

where $s(q,i) = \cos(E_q(q), E_i(i)) \in [-1,1]$ is the cosine similarity between the query encoder output $E_q(q)$ and item encoder output $E_i(i)$, and τ is a learned scalar temperature initialized at 20. The temperature is necessary: without it, $\sigma(s)$ is bounded in $[\sigma(-1), \sigma(1)] \approx [0.27, 0.73]$, preventing the loss from reaching zero for perfect predictions and severely compressing the gradient signal. The temperature-scaled variant is standard in cosine-based contrastive losses [25].

Stage 2: Multiple Negatives Ranking Loss (Discrimination). The model learns to distinguish relevant items from hard negatives ($\mathcal{D}_{\text{HN}}$) that fooled at least one production system. MNR is chosen over BCE because it explicitly optimizes the ranking objective—the model must score the positive higher than *all* in-batch negatives simultaneously:

$$\mathcal{L}_{\text{MNR}} = -\log \frac{\exp(s(q,i^+)/\tau)}{\exp(s(q,i^+)/\tau) + \sum_{j=1}^{n} \exp(s(q,i_j^-)/\tau)} \tag{5}$$

where τ is the same learned temperature as in Stage 1 (carried over via checkpoint transfer). This stage combines hard positives ($\mathcal{D}_{\text{HP}}$, rating ≥ 3) with hard negatives, exposing the model to items the current ANN system retrieves incorrectly. A known limitation of in-batch negative sampling is that another query's positive may be partially relevant to the current query, creating false negatives. We mitigate this through large batch sizes (which reduce per-pair collision probability) and by relying on the subsequent Triplet stage to refine the embedding space; a more principled solution such as debiased contrastive estimation [30] is left for future work.

Stage 3: Triplet Loss (Semantic Nuance). The model learns fine-grained discrimination using token-similar negatives ($\mathcal{D}_{\text{SHN}}$) that share vocabulary with the query but are semantically irrelevant. Triplet loss with margin α directly enforces an embedding-space gap between positive and negative items, which is critical when the negatives are lexically similar and thus close in the initial embedding space:

$$\mathcal{L}_{\text{triplet}} = \max\big(0,\ d(q,i^+) - d(q,i^-) + \alpha\big) \tag{6}$$

where $d(a,b) = 1 - \cos(E(a), E(b))$ is cosine distance (so that lower values indicate higher similarity, consistent with margin semantics) and $\alpha = 0.3$ is the margin hyperparameter selected via grid search on a held-out validation set. This stage forces the model to learn beyond surface-level keyword matching.

Model architecture. The student model is a two-tower architecture based on sentence-transformers [27], initialized from a 6-layer, 22.7M-parameter transformer encoder that produces 384-dimensional embeddings. The query and item encoders *share all transformer layers*—a Siamese configuration that reduces model size and regularizes the shared embedding space. The query encoder receives the raw query string; the item encoder receives the product title. No additional product attributes (description, brand, category) are used in the current version; incorporating structured item features is a direction for future work. Training uses AdamW [19] with learning rate 1×10^{-5}, 5% linear warmup, batch size 512 per GPU across 4 GPUs (effective batch size 2048), mixed-precision (float16), and `torch.compile` optimization. Models are exported to ONNX format and served via a GPU inference server with approximate nearest neighbor search, meeting production latency requirements.

5. Experimental Setup

5.1. Evaluation Dataset

Offline evaluation uses a held-out set of production queries that are *entirely disjoint from the 4M queries used for training data extraction.* No query in the evaluation set appears in any stage of the training pipeline—neither in the multi-channel retrieval collection, the cascade annotation, nor the curriculum training. This separation ensures that offline metrics reflect generalization to unseen queries, not memorization of training examples.

The evaluation queries are stratified by traffic volume. The segment sizes are unequal for practical reasons:

- **Head**: 1,439 queries (highest-frequency searches)—a small set because few queries individually reach head-level traffic, but each is high-impact.
- **Torso**: 21,185 queries (moderate-frequency searches)—the largest segment, reflecting that torso queries are both numerous and individually frequent enough to annotate comprehensively.
- **Tail**: 7,679 queries (low-frequency, long-tail searches)—a sampled subset, because the space of possible tail queries is virtually unbounded and exhaustive coverage is infeasible. We sample tail queries to ensure representation without claiming coverage of the full long-tail distribution.

This stratification means the overall metrics in Table 2 are weighted toward torso queries (the largest segment), while the per-segment breakdown in Table 4 isolates performance on each frequency tier—particularly tail queries, where click-based models are most deficient.

The evaluation set comprises 30,303 unique queries (1,439 + 21,185 + 7,679). For each query, we evaluate the top-500 retrieved items from each system, producing approximately 15.2 million query-item pairs. To ensure a fair comparison on identical query distributions, we restrict evaluation to queries for which both systems return results; in practice, fewer than 2% of queries are unique to one system, so this restriction has negligible effect on the comparison.

Evaluation labels. Relevance labels for evaluation are provided by trained *third-party human annotators*—distinct from and independent of the LLM cascade used to produce training labels. Annotators follow the same 5-class graded-relevance guidelines used in Walmart's production quality measurement, and each query-item pair is labeled by two annotators with a third used for adjudication on disagreements; inter-annotator agreement (Cohen's κ) on a 10K-pair audit is 0.79 (substantial agreement on the 5-class scale), with per-class agreement highest on the extreme ratings (Embarrassing: κ=0.88, Excellent: κ=0.95) and lowest on adjacent middle classes (Okay: κ=0.62, Bad: κ=0.68). Using human labels for the evaluation set means that the offline metrics in Section 6 are not subject to the circularity concern of scoring a model against the same supervision signal that trained it: the training set uses cascade labels, but the evaluation set is fully independent human supervision. The 89.1% cascade-human agreement reported in Section 4.2 is computed on this same human-labeled pool, providing a consistent calibration point across training and evaluation. The online A/B test (Section 5.5) provides a second, annotation-free source of validation through business metrics.

5.2. Baselines

We compare Embedding Model V3 (our system) against two baselines on the same 30K human-labeled query set:

- **Embedding Model V2 (primary baseline)**: the previous-generation two-tower BERT retrieval model deployed on Walmart ANN search [38], trained using click-based labels with a single-stage training procedure. This is the direct head-to-head comparison: same architecture, same serving stack, different training supervision.
- **Pre-trained base model (reference)**: the same transformer encoder [27] that both V2 and V3 are initialized from, evaluated without any fine-tuning. This baseline isolates the effect of fine-tuning: any gap between the base model and V2 reflects what click-based training adds, and any gap between V2 and V3 reflects what our LLM-annotated pipeline adds on top.

The three retrieval channels (Section 4.1) are data sources for training extraction, not evaluation baselines.

5.3. Significance Testing

For each offline metric we report paired bootstrap 95% confidence intervals over the 30K evaluation queries (10,000 resamples), and we test V2-vs-V3 differences with a two-sided paired bootstrap test at $\alpha = 0.05$. For the online A/B test, all reported lifts are verified as statistically significant via the platform's internal experimentation framework. All offline V3-vs-V2 differences reported in Section 6 are significant at $p < 10^{-3}$; 95% CIs are reported in Table 2.

5.4. Metrics

Our primary offline metric is NDCG@10 [16], computed with raw 5-class relevance gains (0–4) as the gain function. We also report Precision (fraction of top-10 items with rel ≥ 3), MAP, MRR, and Average Relevance (mean score on the 0–4 scale). Recall requires special care in our setting: since the full set of relevant items in the catalog is unknown, we compute pool-based recall over the judged pool (union of top-500 retrieved items from both systems for each query). This may underestimate absolute recall for both systems equally, but relative comparisons remain valid.

5.5. Online A/B Test Setup

The online evaluation was conducted on Walmart's sponsored search platform over a two-week period with randomized user-bucket assignment:

- **Control**: Embedding Model V2 (production baseline).
- **Treatment**: Embedding Model V3 (our system).
- **Metrics**: Ad spend, CTR, eCPM, ROAS, click conversion rate, revenue per click.

Each arm received tens of millions of ad requests over the two-week test period, providing high statistical power for detecting the observed effect sizes.

6. Results

We present our results in five parts. First, we report overall offline metrics. Second, we analyze the relevance distribution shift. Third, we break down results by query segment and product category. Fourth, we present online A/B test results. Fifth, we report ablation studies.

Table 2
Overall offline evaluation results (30,303 human-labeled evaluation queries). The pre-trained base model is the same encoder that V2 and V3 are initialized from, evaluated without fine-tuning. Values in brackets are paired bootstrap 95% CIs on the absolute Δ (V3 vs. V2). All V3-vs-V2 differences are significant at $p < 10^{-4}$ (paired bootstrap, 10K resamples).

Metric	Pre-trained base	Model V2	Model V3	V3 vs. V2 (95% CI)
Avg Relevance	2.040	2.413	2.687	+0.274 [+0.264, +0.285]
NDCG@10	0.771	0.878	0.923	+0.045 [+0.042, +0.051]
Precision	0.413	0.526	0.624	+0.098 [+0.093, +0.113]
Recall	0.509	0.676	0.801	+0.125 [+0.121, +0.135]
MAP	0.588	0.722	0.820	+0.098 [+0.095, +0.102]
MRR	0.811	0.901	0.945	+0.044 [+0.038, +0.048]

Table 3
Relevance distribution shift. Values are percentages of all retrieved items at each rating level. The largest absolute change is the 5.2pp reduction in embarrassing results.

Rating	Model V2	Model V3	Change
0 – Embarrassing	8.7%	3.5%	-5.2pp
1 – Bad	26.6%	22.6%	-4.0pp
2 – Okay	12.0%	11.5%	-0.5pp
3 – Good	20.0%	26.6%	$+6.6$pp
4 – Excellent	32.7%	35.8%	$+3.1$pp

6.1. Offline Evaluation

Table 2 presents the overall offline evaluation results. We include the pre-trained base model (the same encoder that V2 and V3 are initialized from, without any fine-tuning) to isolate the contribution of training. The base model achieves NDCG@10 of 0.771, showing that V2's click-based training already provides a substantial +13.9% gain (0.771 $\rightarrow$ 0.878). Our LLM-annotated pipeline pushes this further to 0.923 (+5.1% over V2), indicating that V3's improvements are relative to an already well-trained production system.

6.2. Relevance Distribution Shift

Beyond aggregate metrics, Table 3 reveals the qualitative shift in retrieval quality. Embedding Model V3 reduces the proportion of Embarrassing results (rating 0) from 8.7% to 3.5%—a 5.2 percentage point reduction—while increasing Good and Excellent results from a combined 52.7% to 62.4%. This shift suggests that the LLM-annotated supervision improves average relevance in part by reducing the most harmful retrieval errors.

Figure 2 compares NDCG@10 and precision across query segments, highlighting that improvements are universal but largest for tail queries where click-based training is most deficient.

6.3. Performance by Query Segment

Table 4 shows that improvements are consistent across all query segments, with particularly notable gains for tail queries where the production baseline performs poorest. Tail queries improve from 0.830 to 0.886 NDCG@10 (+6.8%), the largest relative gain among all segments. This is consistent with our hypothesis: cascade annotations do not depend on historical click volume, so tail queries receive the same annotation quality as head queries, closing the data-sparsity gap that degrades click-trained models.

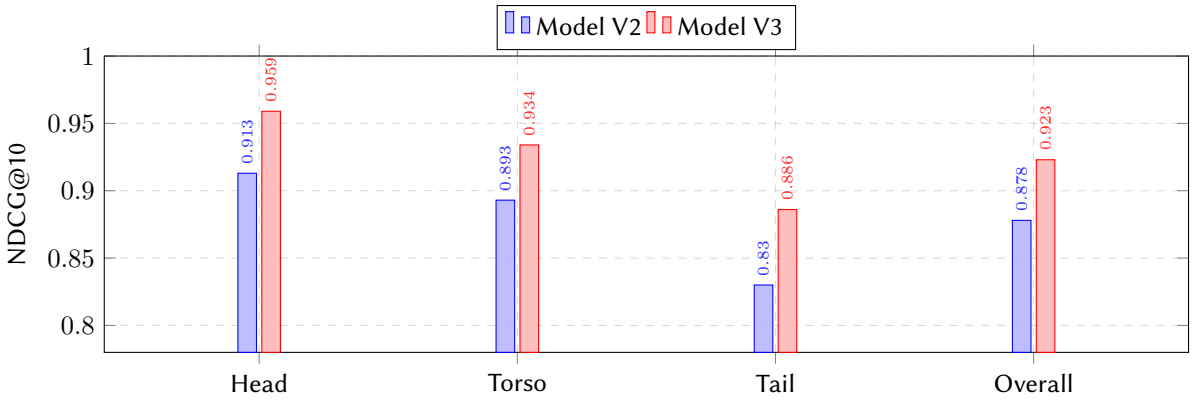

Figure 2: NDCG@10 by query segment. Tail queries show the largest relative improvement (+6.8%), consistent with the hypothesis that LLM-annotated supervision addresses the data-sparsity challenge inherent to click-based training.

Table 4
Performance by query segment (evaluation queries). NDCG@10 and Avg Relevance reported.

Segment	Queries	NDCG@10		Avg Relevance	
		V2	V3 (Δ)	V2	V3 (Δ)
Head	1,439	0.913	0.959 (+5.1%)	2.527	2.812 (+11.3%)
Torso	21,185	0.893	0.934 (+4.6%)	2.469	2.733 (+10.7%)
Tail	7,679	0.830	0.886 (+6.8%)	2.239	2.536 (+13.3%)

Table 5
Performance by product category (evaluation queries, NDCG@10).

Category	Model V2	Model V3	Improvement
Apparel	0.933	0.962	+0.029 (+3.1%)
Consumables	0.894	0.935	+0.041 (+4.5%)
Entertainment	0.883	0.923	+0.040 (+4.5%)
Food	0.837	0.897	+0.060 (+7.1%)
Hardlines	0.905	0.946	+0.041 (+4.5%)
Home	0.925	0.957	+0.032 (+3.4%)

6.4. Performance by Product Category

Table 5 shows consistent improvements across all six major product categories, from high-performing categories (Apparel: +3.1%) to challenging ones (Food: +7.1%, from 0.837 to 0.897 NDCG@10). The consistency of improvements across diverse product categories suggests that the multi-channel training data and LLM annotation generalize well across product domains.

6.5. Online A/B Test Results

Table 6 presents the results of the two-week online A/B test conducted on Walmart's sponsored search platform. With tens of millions of ad requests per arm, the test has high statistical power for the observed effect sizes. Embedding Model V3 achieves consistent improvements across all business metrics, providing annotation-independent evidence that the offline relevance gains translate to real economic value.

The simultaneous improvement in both efficiency metrics (eCPM, CTR) and downstream business metrics (ROAS, conversion, revenue per click) is consistent with a positive feedback loop: more relevant ads may attract more clicks, which improves advertiser ROI and can incentivize higher bids, though we

Table 6
Online A/B test results (two weeks, tens of millions of ad requests per arm, randomized user-bucket assignment). Relative lift of Embedding Model V3 (treatment) over Embedding Model V2 (control). All lifts are statistically significant as verified by the platform's experimentation framework.

Metric	Relative Lift
Ad Spend	+2.80%*
Ad CTR	+1.4%*
eCPM	+2.8%*
ROAS Direct	+2.4%*
Clicks Conv. Direct	+2.9%*
Rev/Click Direct	+3.9%*

* Statistically significant ($p < 0.05$).

Table 7
Ablation study (30,303 human-labeled evaluation queries). Each row removes or modifies one component from the full V3 system (NDCG@10 = 0.923).

Configuration	NDCG@10	Δ vs. V3
V3 (full: BCE → MNR → Triplet)	0.923	—
Training objective (curriculum ordering)		
– single-stage (all data mixed)	0.843	−0.080 (−8.7%)
– MNR → Triplet (drop BCE)	0.903	−0.020 (−2.2%)
– Triplet → MNR (drop BCE)	0.897	−0.026 (−2.8%)
– BCE → MNR (drop Triplet)	0.883	−0.040 (−4.3%)
– BCE → Triplet (drop MNR)	0.857	−0.066 (−7.2%)
– MNR → BCE (drop Triplet)	0.843	−0.080 (−8.7%)
– Triplet → BCE (drop MNR)	0.844	−0.079 (−8.6%)
Training data mining		
– single-channel extraction (ANN only)	0.875	−0.048 (−5.2%)
– without token-similar (semi-hard) negatives	0.886	−0.037 (−4.0%)
– without multi-channel agreement filter	0.916	−0.007 (−0.8%)
– without hard positives (lexical \ ANN)	0.915	−0.008 (−0.9%)
Annotation cascade		
– Platt scaling instead of isotonic	0.904	−0.019 (−2.1%)
– cross-encoder only (no LLM fallback)	0.914	−0.009 (−1.0%)
– global isotonic calibration (no per-class)	0.917	−0.006 (−0.7%)
– 8B LLM only (no routing, ∼3× cost)	0.921	−0.002 (−0.2%)

cannot isolate these causal mechanisms from a single A/B test.

6.6. Component Analysis and Ablations

We isolate the contribution of each major component through a set of controlled ablations on the full human-labeled evaluation set. Table 7 reports NDCG@10 for each ablation, with V3 (full system) at the top as the reference and individual component removals grouped into three categories: training objective, training data mining, and annotation cascade. We discuss each group in turn.

Curriculum training. The three-stage curriculum is the single largest contributor: single-stage training on the same 240M examples scores 0.843, while the full BCE → MNR → Triplet curriculum reaches 0.923 (+9.5%). The six two-stage variants reveal two clear patterns. First, *stage ordering matters enormously*: any configuration ending with BCE collapses to near-single-stage performance (MNR → BCE = 0.843, Triplet → BCE = 0.844). BCE is a foundation loss that teaches coarse relevance discrimination; using it as a refinement stage after a ranking loss destroys the fine-grained structure

the ranking loss learned. Second, *MNR before Triplet is better than Triplet before MNR*: MNR → Triplet = 0.903 vs. Triplet → MNR = 0.897. The listwise ranking objective (MNR) provides a stronger intermediate representation for the pairwise margin objective (Triplet) to refine. The BCE foundation stage adds +2.2% on top of the best two-stage variant (0.903 → 0.923), confirming that the coarse-to-fine progression is genuinely additive rather than redundant.

Training data mining. Multi-channel extraction is the largest data-side contributor: restricting to a single channel (ANN only) costs 5.2% NDCG@10 (0.875 vs. 0.923). Token-similar negatives are the second largest at 4.0% (0.886), confirming that Stage 3's lexically-plausible distractors are essential for teaching the model to look beyond surface keyword matching. The agreement filter and hard positives contribute smaller but consistent gains (0.8% and 0.9% respectively). The agreement filter's modest impact suggests that cross-channel agreement is a useful but not critical signal for negative quality; the bulk of the data mining value comes from channel diversity and the token-similarity mining.

Annotation cascade. The cascade ablations show that calibration method matters more than model selection. Platt scaling is the worst performer (0.904, -2.1%), substantially worse than per-class isotonic (0.923) and global isotonic (0.917, -0.7%). This confirms the value of per-class calibration for heterogeneous confidence distributions. On the model side, using only the cross-encoder costs 1.0% but handles 74.5% of pairs, while using the 8B LLM alone scores 0.921—nearly matching the full cascade at roughly $3\times$ the compute cost. The cascade's value is primarily economic: it achieves equivalent accuracy at a fraction of the annotation budget.

Summary. Ordered by impact: (1) three-stage curriculum with correct ordering (+9.5%), (2) multi-channel extraction (+5.2%), (3) token-similar negatives (+4.0%), (4) per-class isotonic calibration (+2.1% vs. Platt), (5) cross-encoder-to-LLM cascade routing (+1.0%), (6) hard positives (+0.9%), (7) agreement filter (+0.8%), and (8) per-class vs. global isotonic (+0.7%). The first three components account for the majority of the improvement, which aligns with the paper's framing: the structured data mining and curriculum design are the core contributions.

7. Discussion

We discuss why our approach works, what trade-offs it involves, and where it still falls short.

Why does LLM-annotated supervision work? We believe three factors contribute. First, the cascade provides relevance signals that do not depend on click position or query frequency. Unlike click-trained models, which receive no signal for tail queries, the cascade judges every query-item pair on text semantics alone. The cascade has its own biases—inherited from its fine-tuning data—but these biases are systematic and predictable, unlike the heterogeneous noise in click signals. Second, the multi-channel mining pipeline creates training examples at five difficulty levels, so the model sees a diverse range of failure modes: missed relevant items, retrieved irrelevant items, and lexically similar distractors. Third, the curriculum prevents the model from being overwhelmed by hard examples before it learns basic relevance patterns, which the +9.5% NDCG@10 gain over single-stage training confirms empirically. The ablation results further reveal that stage ordering is critical: ending with the coarse BCE loss destroys the fine-grained structure learned by ranking losses, collapsing performance to near-single-stage levels.

Why multi-channel disagreement matters. The low overlap between retrieval channels (13.4% dictionary-ANN, 15.05% dictionary-BM25 at $K{=}500$) is a feature, not a limitation. It means most items fall into regions where systems disagree, and these are exactly the cases that matter most for training. Agreement regions provide high-confidence positives. Disagreement regions provide hard positives (items lexical systems find but the ANN misses) and hard negatives (items one system retrieves incorrectly). These are qualitatively different from random negatives: they are the items that confuse real production systems, making them maximally informative for contrastive training.

Why tail queries improve the most. The largest relative gain is on tail queries (+6.8% NDCG@10), where click-trained models perform worst due to sparse engagement data. This is expected: the cascade assesses relevance based on text, so rare queries and new items receive the same annotation quality as

popular ones. Combined with the multi-channel mining that surfaces hard positives for tail queries, our approach closes the quality gap between query frequency segments.

Hard labels vs. soft labels: a trade-off. We use hard (argmax) cascade labels rather than soft probability distributions. This decouples the cascade from the student entirely: annotations are computed once, cached, and reused across training runs and model architectures. However, it discards the cascade's uncertainty on borderline pairs (ratings 1–3), where collapsing the distribution to a single class may propagate systematic biases. Quantifying this gap through a controlled comparison of hard-label training against soft-label distillation from the same cascade is the most informative follow-up to this work.

Where does the model still fail? The Food category shows the lowest absolute NDCG@10 (0.897) even after improvement. Food queries often involve ambiguous specifications (e.g., "organic milk" vs. specific brands) and high synonym variation. This suggests that the current title-only input could benefit from structured product attributes such as brand, category, and description. The token-similar negative mining may also be less effective for categories where genuinely different products share most of their vocabulary.

7.1. Limitations

We note the following limitations of our work. (a) *Proprietary data*: our training and evaluation data are proprietary to Walmart and cannot be publicly released, though the methodology is described in sufficient detail to be reproduced on other corpora. (b) *Training-label noise*: the 89.1% cascade-human agreement means roughly 11% of training labels diverge from human judgment, with the largest errors on borderline classes (1–3). (c) *Comparison breadth*: our primary comparison is against the production V2 baseline; a broader comparison against fine-tuned general-purpose retrievers would further characterize where gains come from. (d) *Multi-channel dependency*: the pipeline requires multiple heterogeneous retrieval systems; platforms with a single channel would need to create synthetic disagreement. (e) *Single deployment*: the online A/B results come from one two-week test; replication on additional markets and longer time windows would strengthen the evidence.

7.2. Production Considerations

The system runs in production with a weekly retraining cadence. The fully automated pipeline—from multi-channel data extraction through cascade annotation, curriculum training, and ONNX export—eliminates manual annotation and enables continuous model improvement. Serving uses GPU-accelerated inference with approximate nearest neighbor search, meeting production latency requirements. The total retraining cycle completes within 24–48 hours on GPU infrastructure.

A subtlety of weekly retraining is the feedback loop: the ANN channel (R_A) is the current production model, so the hard positives shift each week as the ANN improves. In practice, we observe that this loop is stable—each iteration's hard-positive set shrinks as the ANN closes gaps with the lexical systems. However, we have not formally characterized the convergence properties of this loop, and pathological drift remains a theoretical risk.

8. Conclusion

How can we train dense retrieval models at production scale without click signals or manual annotation? In this paper, we answered this question with an end-to-end pipeline that combines multi-channel retrieval mining, calibrated LLM cascade annotation (89.1% agreement with human annotators), structured sample extraction across five difficulty levels, and three-stage progressive curriculum training. The pipeline generates 240M+ training examples across 4M queries and enables fully automated weekly model retraining.

We deployed the resulting model on Walmart's sponsored search and evaluated it against 30K queries with independent third-party human labels. The system improves NDCG@10 by +5.1% over the click-

trained production baseline, with gains across all query segments and product categories. Notably, the largest gain appears on tail queries (+6.8%), where click-based models have the least training signal. Embarrassing retrievals drop by more than half (8.7% $\rightarrow$ 3.5%). A two-week online A/B test with tens of millions of ad requests per arm provides independent validation: +2.80% ad spend, +2.8% eCPM, and +2.9% click conversion rate.

Overall, our work provides a practical blueprint for replacing click-based training with structured LLM-annotated supervision in production retrieval systems. The key insight—that multi-channel retrieval disagreement is a rich source of structured training signal—is applicable to any platform with multiple retrieval systems. Future work will pursue three directions: (1) enriching the input representation with structured product attributes and multimodal signals, as the current title-only approach underperforms on visually-driven and attribute-heavy categories; (2) comparing hard-label distant supervision against soft-label distillation from the same cascade to quantify the information loss from label discretization; and (3) extending the framework to cross-market and multilingual settings.

Declaration on Generative AI

During the preparation of this work, the author(s) used Generative AI only for formatting assistance in plotting code used to produce figures, and not for drafting, analysis, interpretation, or any other part of the manuscript. The author(s) reviewed, corrected, and validated the generated code and figures and take(s) full responsibility for the publication's content.

References

[1] Luca Maria Aiello, Ioannis Arapakis, Ricardo Baeza-Yates, Xiao Bai, Nicola Barbieri, Amin Mantrach, and Fabrizio Silvestri. The role of relevance in sponsored search. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*, pages 185–194, 2016.

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 41–48, 2009.

[3] Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. InPars: Data augmentation for information retrieval using large language models. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2316–2320, 2022.

[4] Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.

[5] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for YouTube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, 2016.

[6] Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y Zhao, Aida Rashid, Mike Green, and Kelvin Guu. Promptagator: Few-shot dense retrieval from 8 examples. In *International Conference on Learning Representations*, 2023.

[7] Shasvat Desai, Md Omar Faruk Rokon, Jhalak Nilesh Acharya, Isha Shah, Hong Yao, Utkarsh Porwal, and Kuang-chih Lee. Unified supervision for walmarts sponsored search retrieval via joint semantic relevance and behavioral engagement modeling. *arXiv preprint arXiv:2604.07930*, 2026.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019.

[9] Guglielmo Faggioli, Laura Dietz, Charles Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on the Theory of Information Retrieval*, pages 39–50, 2023.

[10] Miao Fan, Jiacheng Guo, Shuai Zhu, Shuo Miao, Mingming Sun, and Ping Li. MOBIUS: Towards the next generation of query-ad matching in Baidu's sponsored search. In *Proceedings of the 25th*

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 2509–2517, 2019.

[11] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292, 2021.

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[13] Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. Improving efficient neural ranking models with cross-architecture knowledge distillation. *arXiv preprint arXiv:2010.02666*, 2020.

[14] Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. Embedding-based retrieval in Facebook search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2553–2561, 2020.

[15] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. Learning deep structured semantic models for web search using clickthrough data. In *ACM International Conference on Information and Knowledge Management (CIKM)*, 2013.

[16] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[17] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, 2020.

[18] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023.

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2019.

[20] Alessandro Magnani, Feng Liu, Suthee Chaidaroon, Sachin Yadav, Praveen Reddy Suram, Ajit Puthenputhussery, Sijie Chen, Min Xie, Anirudh Kashi, Tony Lee, et al. Semantic retrieval at Walmart. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3495–3503, 2022.

[21] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, 2009.

[22] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

[23] Priyanka Nigam, Yiwei Song, Vijai Mohan, Vihan Lakshman, Weitian Ding, Ankit Shingavi, Choon Hui Teo, Hao Gu, and Bing Yin. Semantic product search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2876–2885, 2019.

[24] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 5835–5847, 2021.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[26] Chandan K Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. Shopping queries dataset: A large-scale ESCI benchmark for improving product search. In *Proceedings of the 28th ACM SIGKDD Conference*

on Knowledge Discovery and Data Mining, pages 4429–4439, 2022.

[27] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[28] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[29] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

[30] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2021.

[31] Md Omar Faruk Rokon, Andrei Simion, Weizhi Du, Musen Wen, Hong Yao, and Kuang-chih Lee. Enhancement of e-commerce sponsored search relevancy with LLM. In *Proceedings of the SIGIR Workshop on eCommerce (eCom'24)*, 2024.

[32] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017.

[33] Ning Su, Jiyin He, Yiqun Liu, Min Zhang, and Shaoping Ma. User intent, behaviour, and perceived satisfaction in product search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 547–555, 2018.

[34] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[35] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. Large language models can accurately predict searcher preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1930–1940, 2024.

[36] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

[37] Yuanxing Wang, Yaqing Xue, Buyun Liu, Musen Wen, Wenjia Zhao, and Song Guo. Click-conversion multi-task model with position bias mitigation for sponsored search in ecommerce. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023.

[38] Zhaodong Wang, Weizhi Du, Md Omar Faruk Rokon, Prabir Adhikary, Yaqing Xue, Jian Xu, Jingyi Zhou, Kuang-chih Lee, and Musen Wen. Semantic ads retrieval at Walmart ecommerce with language models progressively trained on multiple knowledge domains. *arXiv preprint arXiv:2502.09089*, 2025.

[39] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-Pack: Packed resources for general Chinese embeddings. *arXiv preprint arXiv:2309.07597*, 2023.

[40] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*, 2021.

[41] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. Optimizing dense retrieval model training with hard negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1503–1512, 2021.

[42] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.