

Embedding Subspace Partitioning for Dynamic Multi-Objective Retrieval

Shaobo Zhang
LinkedIn

Mountain View, CA, USA
shaozhang@linkedin.com

Alice Leung
LinkedIn

Mountain View, CA, USA
alleung@linkedin.com

Yunxiang Ren
LinkedIn

Mountain View, CA, USA
yunren@linkedin.com

Ping Liu
LinkedIn

Mountain View, CA, USA
piliu@linkedin.com

Yuchin Juan
LinkedIn

Mountain View, CA, USA
yjuan@linkedin.com

Qianqi Shen
LinkedIn

Mountain View, CA, USA
qishen@linkedin.com

Benjamin Le
LinkedIn

Mountain View, CA, USA
ble@linkedin.com

Jianqiang Shen
LinkedIn

Mountain View, CA, USA
jshen@linkedin.com

Chengming Jiang
LinkedIn

Mountain View, CA, USA
cjiang@linkedin.com

Ko-Cheng Wang
LinkedIn

Mountain View, CA, USA
lucwang@linkedin.com

Vidya Krishnamurthy
LinkedIn

Mountain View, CA, USA
srkrishnamurthy@linkedin.com

Caleb Johnson
LinkedIn

Mountain View, CA, USA
cajohnson@linkedin.com

Fedor Borisjuk
LinkedIn

Mountain View, CA, USA
fborisyuk@linkedin.com

Luke Simon
LinkedIn

Mountain View, CA, USA
lsimon@linkedin.com

Jingwei Wu
LinkedIn

Mountain View, CA, USA
jingwu@linkedin.com

Wenjing Zhang
LinkedIn

Mountain View, CA, USA
wzhang@linkedin.com

Abstract

Modern industrial recommender systems must optimize across competing objectives, balancing semantic relevance with business metrics such as engagement and revenue. While bi-encoders dominate large-scale retrieval due to their efficiency, they collapse these heterogeneous signals into a single static embedding space. This design creates a fundamental limitation: once trained, the retriever cannot adapt to shifting objective priorities at serving time without retraining. Moreover, joint optimization with multi-objective losses often induces interference between objectives, leading to suboptimal trade-offs. We propose *Embedding Subspace Partitioning (ESP)*, a retrieval framework that decomposes the embedding into task-aware subspaces and replaces the single dot product with a weighted sum of per-subspace similarities, whose weights are tunable at serving time. For Transformer bi-encoders, ESP uses the model’s native end-of-sequence token as a segment delimiter, with segment-aware attention masking and position encoding resets to guarantee subspace isolation in a single forward pass. Serving is performed via GPU-accelerated exhaustive kNN over one concatenated index, eliminating the need for per-objective Approximate Nearest Neighbor (ANN) infrastructure required by multi-head approaches. We evaluate ESP on an open-source benchmark built from MS MARCO [2]. A single ESP model traces a broad Pareto frontier, consistently outperforming strong multi-task baselines across diverse operating points. In LinkedIn’s job matching platform (70M+

weekly users), ESP enabled dynamic retrieval reconfiguration and delivered significant key business metric lifts.

CCS Concepts

- Information systems → Learning to rank; Personalization;
- Computing methodologies → Search methodologies.

Keywords

Retrieval, Multi-Objective Optimization, Embedding, Task-Aware Subspace, Bi-Encoder

ACM Reference Format:

Shaobo Zhang, Alice Leung, Yunxiang Ren, Ping Liu, Yuchin Juan, Qianqi Shen, Benjamin Le, Jianqiang Shen, Chengming Jiang, Ko-Cheng Wang, Vidya Krishnamurthy, Caleb Johnson, Fedor Borisjuk, Luke Simon, Jingwei Wu, and Wenjing Zhang. 2026. Embedding Subspace Partitioning for Dynamic Multi-Objective Retrieval. In *20th ACM Conference on Recommender Systems (RecSys '26)*, September 27–October 02, 2026, Minneapolis, MN, USA. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3773078.3831892>

1 Introduction

Industrial retrievers rarely serve a single notion of relevance. For example, LinkedIn’s job matching platform must produce a ranked candidate set that simultaneously captures semantic fit between members and jobs, predicts member engagement, and aligns with the marketplace’s revenue objectives. Similar multi-objective trade-offs arise across recommendation domains, including e-commerce search (semantic relevance vs. brand and attribute matching) and news feeds (topical relevance vs. freshness and diversity). As a result, the first-stage retriever must generate candidates that balance multiple, and often competing, optimization objectives.



This work is licensed under a Creative Commons Attribution 4.0 International License. RecSys '26, Minneapolis, MN, USA

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2284-4/2026/09

<https://doi.org/10.1145/3773078.3831892>

The prevailing paradigm for large-scale retrieval relies on bi-encoder architectures [10, 13, 22], which project each item into a monolithic d -dimensional embedding space and quantify relevance via a single scalar dot product. This design underpins their efficiency, enabling Approximate Nearest Neighbor (ANN) [10, 11] indexing and low-latency serving at scale [1, 17]. However, the same design imposes a critical limitation in multi-objective settings: every signal must be packed into one geometry. Practitioners typically respond by (a) training one embedding with a weighted multi-objective loss [8, 17, 35], which bakes the trade-off into training and forces a retrain on every priority change, or (b) running a separate retrieval stack per objective and merging downstream, which fragments infrastructure linearly. We call this the *no-knob* problem: the single-embedding bi-encoder gives operators no in-product dial between objectives such as semantic fit and revenue.

We posit that this is fundamentally architectural rather than algorithmic. Building on recent capacity bounds for bi-encoders [30], we show (Section 3) that under conflicting objectives a shared embedding’s per-objective capacity degrades with both the number and severity of conflicts; gradient-surgery methods such as PCGrad [33] mitigate but cannot remove the loss, because the bottleneck is representational, not optimizer-side.

We propose **Embedding Subspace Partitioning (ESP)**: partition the embedding into task-aware subspaces, each responsible for one objective, and replace the single dot product with a weighted sum of per-subspace dot products whose weights are tunable at serving time. For Transformer bi-encoders, we reuse the model’s native end-of-sequence token as a segment delimiter, with entity-first ordering, segment-aware attention masking, and position-ID reset to guarantee subspace isolation in a single forward pass with no additional parameters; for DNN bi-encoders, ESP is realized via lightweight per-objective projection heads on a shared trunk.

At a glance this resembles multi-head architectures common in ranking, but applying multi-head to first-stage *retrieval* is fundamentally harder: standard ANN indices such as IVFPQ [11] cluster documents and approximate distances via residual quantization, which presupposes that the embedding is a coherent geometric object. A concatenation of k subspace vectors is not: it is independent embeddings glued together, with no single neighborhood structure to cluster on. The alternative, k independent ANN indices, scales infrastructure linearly with the number of objectives. ESP sidesteps both via GPU-accelerated exhaustive KNN [12]: a model–hardware co-design that scores the weighted sum as a brute-force scan over one concatenated index, with weights applied on the request side as a serving-time knob. This approach has been deployed in LinkedIn’s job matching platform for two years, serving 70M+ weekly active users; in a 7-day online A/B test, retuning the serving-time weights alone on a single trained model yielded +4.0% revenue, +3.95% job applications, and –8.95% no-fit rate.

Our key contributions include the following:

- **A capacity argument** explaining why a single embedding’s per-objective resolution degrades under conflicting objectives, motivating partitioning at the architecture level rather than at the loss level (Section 3).
- **ESP, an architecture-level realization of subspace partitioning** for both Transformer and DNN bi-encoders, scored

by a single weighted dot product and served by GPU accelerated exhaustive KNN [12]. ESP exposes the per-objective weights as a serving-time knob, letting operators trace the full Pareto frontier between objectives from a single trained model (Section 4).

- **Empirical validation** on a multi-objective MS MARCO [2] benchmark with hard entity negatives (released open-source) and on LinkedIn’s production system. ESP outperforms multi-loss and gradient-surgery baselines at every matched operating point and delivers the production gains above (Sections 5 and 6).

2 Related Work

Bi-Encoder Retrieval. Dense retrieval has evolved from early bi-encoder models such as DPR [13] and ANCE [31] to more recent large-scale text encoders including E5 [28] and GTE [18]. While these models achieve strong performance for single-objective retrieval, they fundamentally produce a single embedding per input. Existing approaches to multi-objective retrieval therefore either modify the training objective (discussed below) or alter the embedding architecture itself, which is the direction taken by ESP. Recent theoretical work by Weller et al. [30] establishes dimension-dependent capacity bounds for single-objective bi-encoders; in Section 3, we extend this analysis to the multi-objective setting.

Multi-Objective Optimization in Retrieval. Prior work falls into three groups, all of which optimize *training* while leaving the serving-time embedding architecture unchanged.

Multi-Task Single-Head (MTSH and other single-vector variants). Uni-Retriever [35] and Taobao’s system [17] train one embedding under a weighted multi-objective loss; Baidu Search [8] compresses objectives into scalar scores before ANN; JRC [25] balances recall and calibration via dual member embeddings. Multi-Aspect Dense Retrieval (MADR) [15] supervises per-aspect representations in intermediate layers but *fuses them into a single output vector* via a learned aggregator, preserving ANN compatibility at the cost of serving-time knob. These methods collapse the trade-off into training-time choices, produce a single served embedding, and suffer geometric interference that loss weighting cannot eliminate.

Multi-Task Multi-Head (MTMH). MTMH [34] adds k task-specific projection heads on a shared encoder trunk, each producing a separate embedding. At serving time, each head requires its own ANN index with a post-retrieval merge, scaling infrastructure linearly with the number of objectives. ESP shares the partitioning principle but, as discussed in Section 1, sidesteps the per-objective ANN problem via exhaustive KNN over one concatenated index [12], and adds Transformer-specific mechanisms.

Gradient conflict management. PCGrad [33] projects conflicting task gradients during multi-task training, reducing interference but still producing a single embedding without a serving-time knob.

Other Multi-Task and Multi-Vector Approaches. Matryoshka Representation Learning (MRL) [16] partitions embeddings by *resolution*—truncating trailing dimensions for efficiency—rather than by objective. Instruction-based methods such as LLM Embedder [36] and Retrieval-GRPO [19] switch tasks via prompt or RL but produce one embedding per request, offering task selection rather than

tunable trade-offs. OmniSearchSage [1] and Maillard et al. [20] apply multi-task losses without changing the embedding geometry. ColBERT [14] achieves multi-aspect matching through per-token late interaction, at $O(L)$ vectors per document; ESP keeps storage at $k=2-3$ vectors, comparable to standard bi-encoders.

3 Geometric Analysis of Multi-Objective Embedding

We analyze the minimum embedding dimension required for a single bi-encoder to serve k retrieval objectives simultaneously, and show that this minimum grows with a geometric notion of *conflict* between objectives. While previous work [30] focuses on the cardinality of distinct rankings in $\mathbb{R}^d$, we frame the multi-objective problem as one of subspace approximation. We derive a rigorous lower bound that the partitioned construction of ESP saturates in the worst case, providing an architectural guarantee of capacity.

3.1 Problem Formulation

Definition 1 (Multi-Objective Bi-Encoder Retrieval). Given a query set $\mathcal{Q}$, a document set $\mathcal{D}$, and k relevance matrices $R_1, \dots, R_k \in \mathbb{R}^{|\mathcal{Q}| \times |\mathcal{D}|}$, find encoders $f : \mathcal{Q} \rightarrow \mathbb{R}^d$ and $g : \mathcal{D} \rightarrow \mathbb{R}^d$ such that $S = f(\mathcal{Q})g(\mathcal{D})^\top$ approximates each R_i within relative Frobenius error δ :

$$\|S - R_i\|_F \leq \delta \|R_i\|_F \quad \forall i \in [k].$$

Since $\text{rank}(S) \leq d$, the fundamental constraint is determining the smallest d for which a single rank- d matrix can simultaneously δ -approximate the dominant subspaces of all k objectives.

3.2 Capacity Lower Bound for Shared Embeddings

For each $R_i = U_i \Sigma_i V_i^\top$, let the δ -effective right subspace $V_i^\delta \subseteq \mathbb{R}^{|\mathcal{D}|}$ be the span of the smallest set of right singular vectors capturing $1 - \delta^2$ of the squared Frobenius norm. Let $r_i = \dim V_i^\delta$ denote the δ -effective rank of objective i .

Definition 2 (Subspace Conflict). The pairwise conflict between objectives i and j is

$$\varepsilon_{ij} := 1 - \frac{\|P_{V_i^\delta} P_{V_j^\delta}\|_F^2}{\min(r_i, r_j)},$$

where P_V is the orthogonal projector onto V and $\|P_{V_i^\delta} P_{V_j^\delta}\|_F^2 = \sum_l \cos^2 \theta_l$ is the sum of squared cosines of the principal angles $\{\theta_l\}$ between the two subspaces. Hence $\varepsilon_{ij} \in [0, 1]$: $\varepsilon_{ij} = 0$ when one effective subspace is contained in the other (full alignment), and $\varepsilon_{ij} = 1$ when the subspaces are orthogonal.

This metric captures conflict as an intrinsic representational property of the data geometry. It measures the principal-angle distance between the dominant ranking manifolds of two objectives, quantifying how much of one objective’s signal already lies in the other’s geometry, independent of optimizer or loss.

THEOREM 3 (MULTI-OBJECTIVE CAPACITY LOWER BOUND). Any rank- d score matrix S satisfying $\|S - R_i\|_F \leq \delta \|R_i\|_F$ for all $i \in [k]$

must satisfy

$$d \geq \dim \left(\sum_{i=1}^k V_i^\delta \right).$$

Under the symmetric model where $r_i = r$ and $\varepsilon_{ij} = \varepsilon$ for all $i \neq j$, this simplifies to

$$d \geq r \cdot (1 + (k-1)\varepsilon) \cdot (1 - O(\delta)).$$

Equivalently, for a fixed budget d the per-objective δ -effective rank that can be jointly preserved is bounded above by

$$r_{\text{eff}}(d, k, \varepsilon) \leq \frac{d}{1 + (k-1)\varepsilon}.$$

PROOF SKETCH. Let P be the orthogonal projector onto $\text{row}(S)^\perp$. Since $SP = 0$,

$$\|R_i P\|_F = \|(R_i - S)P\|_F \leq \|R_i - S\|_F \leq \delta \|R_i\|_F,$$

so the component of R_i orthogonal to $\text{row}(S)$ captures at most $\delta^2 \|R_i\|_F^2$. Combined with the definition of the δ -effective subspace and Wedin’s perturbation theorem [29], this forces $V_i^\delta \subseteq \text{row}(S)$ up to an $O(\delta)$ angle. Hence $\text{row}(S) \supseteq \sum_i V_i^\delta$, so $d = \text{rank}(S) \geq \dim(\sum_i V_i^\delta)$.

For the symmetric simplification, decompose each V_i^δ as a *shared core* of dimension $r(1 - \varepsilon)$ aligned across all objectives plus a *private complement* of dimension $r\varepsilon$ orthogonal across objectives. Then $\dim(\sum_i V_i^\delta) = r(1 - \varepsilon) + k \cdot r\varepsilon = r(1 + (k-1)\varepsilon)$, and the bound on r_{eff} follows by inversion. $\square$

This is the multi-objective analogue of the single-objective rank bound: representing k potentially-conflicting objectives requires the embedding to span the union of their effective subspaces, which grows linearly in both k and the geometric conflict ε .

3.3 Constructive Upper Bound via Partitioning

THEOREM 4 (PARTITIONING SATURATES THE WORST-CASE BOUND). By partitioning the embedding into k orthogonal blocks of sizes $\{d_i\}$ where $d_i \geq r_i$ and $\sum_i d_i = d$, and dedicating the i -th block to R_i , the resulting score matrix satisfies $\|S - R_i\|_F \leq \delta \|R_i\|_F$ for all i at total dimension $d = \sum_i r_i$. Importantly, this is independent of the conflict pattern $\{\varepsilon_{ij}\}$.

PROOF. Each block is a rank- d_i approximator of R_i alone. By Eckart–Young [6], $d_i \geq r_i$ guarantees δ -accurate approximation within that block, and orthogonality of blocks ensures objectives do not interfere across the partition. $\square$

Implication for ESP. The advantage of ESP is clear based on Theorems 3 and 4: a shared embedding’s required dimension $r(1 + (k-1)\varepsilon)$ grows with the conflict ε , while partitioning’s requirement $\sum_i r_i$ remains constant. The two bounds coincide only in the worst case ($\varepsilon \rightarrow 1$). At intermediate conflicts a shared embedding can in principle attain the smaller bound, but only if training successfully discovers and aligns the data’s shared and private components. Partitioning trades this potential dimensional efficiency for an *architectural guarantee*: the decomposition is built in, not learned. ESP’s per-subspace structure (Section 4) is the direct realization of this guarantee, ensuring each objective receives a conflict-free slice of representation capacity, regardless of how the data’s relevance subspaces overlap. Our experiments (Section 5) confirm

that shared training in practice does not close the gap even with gradient-surgery methods such as PCGrad, leaving substantial capacity unused.

3.4 Estimating Conflict in Practice

Directly computing ε_{ij} requires the SVD of each R_i , which is computationally prohibitive at industrial scale. We use Kendall's τ between rankings induced by each objective on a held-out query sample as a tractable proxy:

$$\hat{\varepsilon}_{ij} = 1 - \tau(\text{rank}_i, \text{rank}_j).$$

Concretely, τ is computed as the average per-query Kendall correlation between rankings induced by R_i and R_j on the eval set's 1,359 queries over the $\sim 39\text{K}$ -passage corpus (Section 5.3). The proxy is *directionally aligned* with subspace conflict: when subspaces overlap heavily, the dominant relevance directions agree and rankings track closely ($\tau \rightarrow 1$, $\hat{\varepsilon} \rightarrow 0$); when relevance subspaces orthogonalize, ranking correlations vanish ($\tau \rightarrow 0$, $\hat{\varepsilon} \rightarrow 1$). The relationship is not a strict equality: singular-value differences can induce ranking disagreement even at $\varepsilon = 0$. We treat $\hat{\varepsilon}$ as an order-of-magnitude indicator throughout the empirical sections.

3.5 Optimal Dimension Allocation

When the budget d falls short of $\sum_i r_i$, partitioning cannot fully accommodate all objectives and dimensions must be allocated.

THEOREM 5 (ALLOCATION BY EFFECTIVE RANK). *Subject to $\sum_i d_i = d$ and $d_i \geq 1$, the allocation that maximizes the worst-case Frobenius coverage $\min_i \|P_{d_i}(R_i)\|_F^2 / \|R_i\|_F^2$ under uniform spectral decay across objectives is*

$$d_i^* = d \cdot \frac{r_i}{\sum_{j=1}^k r_j}.$$

PROOF SKETCH. Truncating R_i at rank d_i leaves residual $\sum_{l>d_i} \sigma_l^2$. Under uniform spectral decay $\sigma_l^2 / \|R_i\|_F^2 \approx \frac{1}{r_i} \mathbb{1}[l \leq r_i]$, the residual is $1 - d_i/r_i$, and worst-case coverage is maximized when each objective's truncation level d_i/r_i is equalized and gives the proportional allocation. $\square$

Resolution gap in practice. Different objectives can have vastly different effective ranks. A *low-resolution* objective such as categorical entity matching has small r_i : with ~ 780 unique entities, sphere-packing implies $r_i = O(\log 780) \approx 10$ dimensions suffice. A *high-resolution* ranking objective has large r_i scaling with corpus complexity. In our experiments (Section 5.5), entity matching saturates at $d_{\text{ent}} = 32$ (3% of $H = 1024$) while semantic quality requires $d_{\text{sem}} \gtrsim 750$. The resulting $\sim 1:20$ resolution gap has an important systems implication: auxiliary business objectives can often be allocated compact subspaces, leaving most dimensions available for high-fidelity semantic retrieval. Objectives involving highly confusable attributes (e.g., near-duplicate product SKUs) require higher resolution and would correspondingly consume larger subspaces.

4 Method: Embedding Subspace Partitioning

Motivated by the geometric analysis of multi-objective retrieval, we introduce Embedding Subspace Partitioning (ESP). ESP structures the embedding space into discrete, task-aware subspaces

that enable controllable, per-objective scoring during inference. ESP is architecture-agnostic; we present two instantiations: (1) for Transformer-based ESP, using the internal representational capacity of autoregression via specialized attention masking; (2) for DNN-based ESP, using per-objective projection heads on a shared trunk. The variant is deployed in LinkedIn's production retrieval system.

The two instantiations are evaluated on different objective pairs: semantic vs. entity-attribute matching for Transformer-ESP (Section 5, on an open-source MS MARCO benchmark [2]) and engagement vs. revenue for DNN-ESP (Section 6, in production). This reflects the labels naturally available in each setting. Both are instances of the same multi-objective retrieval problem, and the partitioning principle is agnostic to which objectives are partitioned.

4.1 ESP Architecture

ESP partitions the embedding into k orthogonal subspaces, with the i -th subspace producing a sub-embedding $\mathbf{e}_i \in \mathbb{R}^{d_i}$ dedicated to the i -th objective. The full embedding is the concatenation $\mathbf{e} = [\mathbf{e}_1; \dots; \mathbf{e}_k] \in \mathbb{R}^d$ with $d = \sum_i d_i$. Each $\mathbf{e}_i$ is L2-normalized so that $\langle \mathbf{e}_i, \mathbf{e}_i' \rangle \in [-1, 1]$; per-objective dimensions $\{d_i\}$ are allocated proportionally to per-objective effective rank, following Theorem 5.

Scoring with a serving-time knob. The retrieval score between query $\mathbf{q}$ and document $\mathbf{d}$ is a weighted sum of per-subspace dot products:

$$S(\mathbf{q}, \mathbf{d}) = \sum_{i=1}^k w_i \cdot \langle \mathbf{e}_i, \mathbf{e}_i' \rangle. \quad (1)$$

The weights $\{w_i\}$ are applied on the request side at *serving time*, not baked into the model: a single trained ESP model can dynamically trace the full Pareto frontier between objectives by varying $\{w_i\}$, with no retraining required. When $w_i = 1$ for all i , Equation 1 reduces to a standard dot product over the concatenated embedding $\mathbf{e}$, preserving full compatibility with existing exhaustive-KNN retrieval infrastructure [12].

Per-subspace training. Each subspace is trained with its own per-objective loss; the choice of loss family is unconstrained by the architecture. Contrastive losses (e.g. InfoNCE for ranking objectives), regression losses (e.g. MSE for revenue prediction), and classification losses (e.g. BCE for binary fitness) can be combined freely across subspaces. LinkedIn's production system (Section 6) uses all three. For ranking tasks, we utilize the InfoNCE loss:

$$\mathcal{L}_i = -\frac{1}{N} \sum_{n=1}^N \log \frac{\exp(\langle \mathbf{e}_i^{(n)}, \mathbf{e}_i'^{(n)} \rangle / \tau)}{\sum_{j=1}^B \exp(\langle \mathbf{e}_i^{(n)}, \mathbf{e}_i'^{(j)} \rangle / \tau)}, \quad (2)$$

where τ is the temperature parameter and B is the batch size. The total training loss is

$$\mathcal{L} = \sum_{i=1}^k \alpha_i \cdot \mathcal{L}_i. \quad (3)$$

Because $\mathcal{L}_i$ depends only on $\mathbf{e}_i$, we have $\partial \mathcal{L}_i / \partial \mathbf{e}_j = 0$ for $j \neq i$: each loss updates only its own subspace's representation direction without gradient interference, so no "compromise direction" between objectives can arise at the output. The shared encoder parameters still receive mixed gradients from all $\mathcal{L}_i$, but the capacity guarantee of Theorem 4 rests on the output partition, not on

the trunk; training-time weights $\{\alpha_i\}$ therefore affect convergence speed rather than geometric capacity.

4.2 Realizing the Subspace Partition

The architecture above leaves open one engineering question: how are the per-subspace embeddings $\{e_i\}$ computed for a given encoder type? We describe two realizations that produce the structurally orthogonal output partition required by Theorem 4.

4.2.1 Transformer Bi-Encoders: EOS-Delimited Segments.

Input formatting. We use the model’s *native end-of-sequence token* (EOS) as a segment delimiter, rather than adding new vocabulary tokens. This is critical: a pretrained transformer encoders already knows how to produce summary representations at EOS positions, so ESP works without fine-tuning. The attribute segment is placed before the main text:

```
<attribute value> <EOS>
Instruct: ... query: <text> <EOS>
```

Documents follow an analogous format. Attribute-first ordering is preferred for two reasons. (i) Under causal attention alone, the regime used in our pretrained model evaluation without encoder modification, the attribute EOS token naturally attends only to attribute tokens because no passage tokens precede it. This provides isolation for the attribute subspace without requiring additional masking. (ii) When combined with the segment-aware mask below, the partition becomes *symmetric*: both EOS embeddings depend only on their own segments, and Theorem 4’s orthogonality assumption holds exactly.

The attribute value on both query and document sides must refer to the *same typed field*, e.g. both contain the person name, or both contain the job title. This assumes an upstream entity tagger produces the attribute, which is standard in production retrieval systems. For queries without a tagged attribute, the attribute segment is omitted and ESP becomes standard single-embedding retrieval.

Segment-aware attention masking. We apply a segment-aware causal attention mask in which each EOS token at position p_i attends only to tokens in its own segment, i.e. positions $[p_{i-1} + 1, \dots, p_i]$. Both EOS embeddings are therefore restricted to their respective segments: the attribute EOS attends only to attribute tokens, while the semantic EOS attends only to passage tokens. This yields symmetric subspace isolation and satisfies the orthogonal block assumption of Theorem 4. Layer normalization, which aggregates across attended positions, also operates within-segment under this mask, so cross-segment leakage cannot occur through any pathway. We use SDPA (not flash attention) [23, 24] to support custom 4D masks.

Position-ID reset (RoPE correction). Modern Transformers such as frontier LLMs use Rotary Position Embeddings (RoPE) [26], which makes hidden states position-dependent. Without correction, the same attribute text at different absolute positions (due to varying passage lengths) produces different embeddings. We fix this by resetting position IDs at each segment boundary: segment i always starts at position 0 regardless of prior segment lengths. This ensures that identical attribute text on the query and document sides

receives identical position encodings, producing cosine similarity ≈ 1 as expected.

Embedding extraction. The hidden state $h_i \in \mathbb{R}^H$ at each segment’s EOS position is as the pre-normalization sub-embedding. If dimension allocation is applied, h_i is truncated to its first d_i dimensions ($\sum_i d_i \leq H$) before L2-normalization to give e_i . *The Transformer realization introduces no additional parameters* beyond the base encoder; the DNN realization (Section 4.2.2) adds k light-weight projection heads.

Ablation summary. Entity-first ordering provides causal isolation for free: identical entity text yields cosine ≈ 0.97 regardless of passage content. Adding the segment-aware mask further improves entity discriminability by +7.2pp (92.8% $\rightarrow$ 100%). Position reset corrects RoPE bias, restoring cosine from ≈ 0.50 to ≈ 0.73 –1.0. We recommend causal-only ESP for pretrained evaluation, adding mask + position reset when fine-tuning.

4.2.2 DNN Bi-Encoders: Projection Heads. For DNN bi-encoders without causal attention, EOS-based segmentation does not apply; we instantiate the partition through k trainable projection heads on a shared encoder trunk. Each head h_i projects the trunk output into an objective-specific sub-embedding: $e_i^q = h_i(f(q))$, $e_i^d = h_i(g(d))$. Independent heads structurally enforce the output partition. LinkedIn’s production retrieval system [12] uses this design with $k = 3$ objectives (engagement, revenue, fitness); see Section 6 for results.

5 Open-Source Experiments

We design our experiments to validate the theoretical framework established in Section 3: we measure geometric conflict ε from real-world data, predict capacity loss, train models with and without ESP, and demonstrate that ESP overcomes the interference ceiling predicted by Theorem 3.

5.1 Benchmark Construction

Standard retrieval benchmarks often conflate semantic relevance with keyword matching. To isolate these objectives, we construct a multi-objective benchmark from MS MARCO [2] specifically designed with hard entity negatives.

Entity Extraction. For each query, we extract multi-word named entities (e.g., “Martin Luther King,” “New York City,” “machine learning”) that appear verbatim in the relevant passage. Generic phrases are filtered.

Hard Negatives. For each query-entity pair, we sample passages from *other queries* that also contain the entity string but answer a different question. For example, a query about “Martin Luther King speeches” gets a hard negative about “Martin Luther King birthday” — same entity, different information need. This creates genuine conflict: a pure entity-matching signal would rank both equally, whereas a semantic signal must distinguish between them.

Dataset Statistics. The benchmark contains 1,993 unique queries split into 13,951 training and 2,093 evaluation (query, positive, hard-negative) examples; the 2,093 eval examples cover 1,359 unique eval queries. Of the eval pairs, **72% are hard negatives** (both

Table 1: Pretrained retrieval (0.6B, 39K corpus, 1,359 queries). Sem: NDCG@10. Ent: Entity Recall@10. Storage: number of vectors per document.

Method	Sem	Ent	Vectors / Doc	Scoring
Baseline (no entity)	.668	.496	1	dot
Entity in prompt	.762	.594	1	dot
MTMH [†] [34]	.677	.833	2	2× dot, merge
ColBERT [14]	.849	.935	L token vectors	MaxSim
ESP (ours)	.843	.932	1	dot

[†]Pretrained MTMH simulated via per-objective prompts.

positive and negative contain the entity) and 28% are easy negatives (entity absent from negative). All retrieval results in Section 5.3 are reported on the 1,359 eval queries. The entity-first input format (Section 4.2.1) is used for all methods.

5.2 Empirical Measurement of Conflict

We first quantify the objective conflict using a pretrained open-source 0.6B embedding model ($H=1024$) [37]. By Kendall’s τ between the semantic ranking (which passage is semantically relevant?) and the entity ranking (which passage contains the entity?), we find:

$$\tau = 0.34, \quad \varepsilon = 1 - \tau = 0.66$$

Applying Theorem 3 for $k=2$ objectives, we predict a maximum effective rank of $r_{\text{eff}} \leq 1024/(1.66) \approx 616$. This represents a 40% reduction in available representational capacity for any shared-embedding model attempting to satisfy both objectives simultaneously. This is the worst-case theoretical limit. The next two subsections verify the prediction empirically: training (Section 5.3) realizes a smaller but monotonic capacity loss in shared embeddings that no choice of loss weighting or gradient surgery can eliminate, while ESP’s partitioned architecture avoids the loss entirely.

5.3 Retrieval Evaluation

We evaluate all methods using Semantic NDCG@10 (MS MARCO relevance) and Entity Recall@10 (fraction of entity-matched passages in top-10) over a corpus of $\sim 39\text{K}$ passages (eval set plus 35K randomly sampled MS MARCO distractors). For each of 1,359 queries, all methods rank the full corpus.

Pretrained baselines (Table 1). We compare four retrieval architectures on the pretrained 0.6B model [37]. Because pretrained MTMH lacks trained projection heads, we approximate per-head specialization by issuing separate prompts for each objective to the same encoder, retrieving the top- $n/2$ results from each corresponding index, and merging the candidate sets downstream.

Without entity information, the baseline achieves .668 semantic NDCG. Adding entity text to the input via instruction-based routing [36] (prepending entity to both query and passage) improves to .762/+14pp. However this is *task switching*, not multi-objective optimization: the single embedding entangles entity and semantic signals with no mechanism for trade-off. MTMH’s $n/2$ merge splits the retrieval budget between objectives, degrading both metrics. ColBERT achieves the strongest overall quality, but at substantially

Table 2: Trained models (0.6B, 1 epoch, 39K corpus, 1,359 queries). MTSH: multi-task single-head with entity weight α_e . ESP weight w_e is tuned at serving time (no retraining). $\pm$ indicates std across queries.

Method	Sem NDCG@10	Ent Recall@10
Baseline (no entity)	.870 $\pm$.220	.679
MTSH ($\alpha_e=1$)	.924 $\pm$.164	.887
MTSH ($\alpha_e=5$)	.875 $\pm$.208	1.000
MTSH ($\alpha_e=10$)	.823 $\pm$.240	1.000
MTSH + PCGrad [33]	.879 $\pm$.205	1.000
ESP ($w_e=0.3$)	.899 $\pm$.188	.908
ESP ($w_e=0.5$)	.900 $\pm$.183	.986
ESP ($w_e=0.8$)	.900 $\pm$.183	.999

higher serving cost. It stores $O(L)$ token vectors per document and performs per token MaxSim interactions instead of a single vector similarity computation, where L denotes document length. In contrast, ESP achieves entity recall comparable to ColBERT at standard bi-encoder cost: it stores a single concatenated embedding per document composed of k subspaces and computes relevance using one weighted similarity function, served by GPU-accelerated exhaustive KNN [12]. ESP’s combined Sem NDCG (.843) exceeds entity-in-prompt (.762) because ESP’s subspace isolation prevents entity signal from interfering with semantic ranking.

Trained interference ceiling (Table 2). Trained ColBERT is omitted because its $\sim L$ -vectors-per-document storage ($\sim 350\text{TB}$) precludes deployment in our target setting; the trained comparison focuses on single-vector and partitioned architectures. Increasing entity emphasis in a single embedding (MTSH) progressively degrades semantic quality while entity recall saturates. PCGrad [33] gradient surgery *mitigates* the interference (-4.6pp semantic vs. -4.9pp for MTSH $\alpha_e=5$) but cannot eliminate it, confirming the bottleneck is *geometric* (embedding capacity), not gradient conflict. ESP breaks through: at $w_e=0.8$, it achieves .900 semantic *and* .999 entity — outperforming the best MTSH at matched entity recall by +2.5pp semantic NDCG ($p < 0.001$, bootstrap 95% CI [+1.6, +3.5] over 1,359 queries), without requiring retraining.

5.4 Pareto Frontier: A Single-Model Solution

Beyond capacity loss, shared embeddings suffer a practical limitation: they provide **no mechanism to control the trade-off** between objectives at serving time, and any emphasis must be baked into training loss weights. ESP resolves this with serving-time control. Table 3 compares ESP’s weight sweep (one trained model) against MTSH’s Pareto frontier (each point requires a separate training run).

ESP traces a smooth Pareto frontier using a single trained model. At comparable levels of entity recall, ESP consistently matches or exceeds the NDCG achieved by MTSH, while MTSH requires retraining for each operating point. The value of ESP lies not in any single operating point but in the ability to trace the Pareto frontier from a single trained model, a capability MTSH lacks.

Table 3: Pareto comparison (1,359 queries): ESP (one model, vary w_e at serving time) vs. MTSH (retrain for each α_e). Rows ordered by increasing entity emphasis.

ESP (one model)		MTSH (retrain each)	
Sem	Ent	Sem	Ent
.870 ($w_e=0$)	.693	.870 [†]	.679
.899 ($w_e=0.3$)	.908	.924 ($\alpha_e=1$)	.887
.900 ($w_e=0.5$)	.986	.875 ($\alpha_e=5$)	1.000
.900 ($w_e=0.8$)	.999	.823 ($\alpha_e=10$)	1.000

[†]Baseline model (no entity in training).

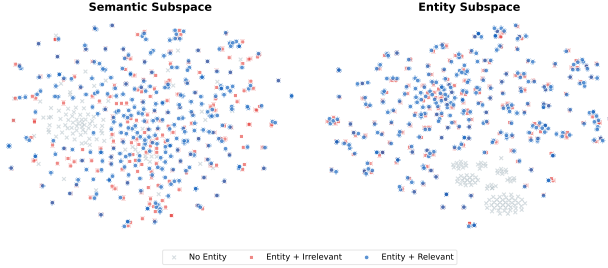


Figure 1: t-SNE of ESP subspace embeddings. Left: Semantic subspace — entity presence is entangled with content. Right: Entity subspace — same-entity passages cluster tightly regardless of relevance.

Semantic NDCG plateau and signal complementarity. ESP’s semantic NDCG is notably stable across $w_e \in [0.3, 0.8]$ (.899 to .900). This plateau reflects *signal complementarity*, not absence of conflict. At $w_e=0$ (semantic only), NDCG is .870; adding entity weight to $w_e=0.3$ improves semantic NDCG by +2.9pp, because entity-matched passages in our benchmark are correlated with semantic relevance. Within the operating range $[0.3, 0.8]$, the passages promoted by increasing w_e are entity-matched and semantically relevant, so semantic quality is maintained. At the extreme ($w_e=1.0$), semantic NDCG collapses to .583, confirming the trade-off exists but only manifests when entity signal overwhelms semantics entirely. This plateau is precisely what subspace isolation enables: because semantic and entity scores are computed independently, entity emphasis cannot *corrupt* the semantic ranking — it can only *reweight* the combined score. In MTSH, where both signals share one embedding, increasing entity emphasis degrades semantic quality monotonically ($0.924 \rightarrow 0.823$) because the signals interfere geometrically.

5.5 Dimension Allocation

We sweep d_{ent} from 0 to H by truncating the hidden state at each EOS, keeping $d_{\text{sem}} = H - d_{\text{ent}}$. With entity-first format providing perfect subspace isolation (cosine ≈ 1.0 for identical entities):

Entity matching requires trivial dimensionality. Both pretrained and trained models achieve 100% entity accuracy at just $d_{\text{ent}} = 32$ (3% of $H=1024$), matching the resolution-gap argument of Section 3.5: with ~ 780 unique entities, $O(\log 780) \approx 10$ bits are needed — well within 32 dimensions.

Table 4: $k=3$ ESP: semantic + entity + length preference (pre-trained 0.6B, 1,359 queries). Sem: Semantic NDCG@10. Ent: Entity Recall@10. Len: Length Match@10 (fraction of top-10 matching preferred length).

Weights (w_s, w_e, w_l)	Sem	Ent	Len
(1.0, 0.0, 0.0) sem only	.867	.687	.812
(0.5, 0.5, 0.0) $k=2$	.861	.985	.858
(0.5, 0.3, 0.2) $k=3$ balanced	.843	.957	.912
(0.4, 0.4, 0.2) $k=3$ ent+len	.824	.969	.912
(0.5, 0.2, 0.3) $k=3$ heavy len	.830	.918	.931

Figure 1 visualizes the subspace separation: entity identity is entangled with semantic content in the shared embedding (left), but cleanly separable in its own subspace (right), confirming entity matching is a low-dimensional signal trivial to isolate.

5.6 Single-Objective Generality

A multi-objective architecture is only useful if it does not regress on single-objective retrieval. We verify ESP’s behavior in this regime by evaluating the pretrained model (with segment mask and position reset) on three BEIR benchmarks [27] using a “none” entity prefix—reducing ESP to a standard bi-encoder. ESP shows $\leq 0.3\text{pp}$ variation vs. baseline across NFCorpus, SciFact, and FiQA, indicating no meaningful regression: the entity subspace embedding becomes constant across all documents and does not influence ranking.

5.7 Scaling to Three Objectives ($k = 3$)

To test ESP’s generality beyond $k=2$, we add a third objective: *passage length preference*. We bucket passages into short (<30 words), medium (30–80), and long (>80), and set the query preference to “medium” (the most informative length). The input format becomes: $\langle \text{length} \rangle \langle \text{EOS} \rangle \langle \text{entity} \rangle \langle \text{EOS} \rangle \langle \text{passage} \rangle \langle \text{EOS} \rangle$, producing three subspace embeddings scored by $S = w_s \cdot s_{\text{sem}} + w_e \cdot s_{\text{ent}} + w_l \cdot s_{\text{len}}$.

We use length preference as a *toy third objective* to demonstrate that ESP’s serving-time control extends to $k > 2$; the production system in Section 6 validates $k=3$ on three high-stakes objectives (engagement, revenue, fitness). At (0.5, 0.3, 0.2), the model achieves .843 semantic, .957 entity, and .912 length match to simultaneously control three objectives from a single pretrained model with no additional training. Compared to $k=2$ at (0.5, 0.5, 0), adding the length objective costs -1.8pp semantic and -2.8pp entity while gaining $+5.4\text{pp}$ length match. This is a smooth, predictable trade-off rather than catastrophic interference.

6 Production Validation at LinkedIn

While the preceding experiments on Transformer-based architectures validate the geometric principles and capacity benefits of ESP, we now demonstrate the framework’s industrial viability and scalability. We deployed ESP within LinkedIn’s job matching platform, a high-throughput system serving over 70M weekly active users. It utilizes the DNN-ESP variant, employing $k=3$ dedicated projection heads branching from a shared encoder trunk. By transitioning from the controlled benchmark to a live production environment,

Table 5: Production multi-objective co-training results. MTSH: single embedding, weighted training loss (α : revenue weight). ESP: partitioned embedding with per-objective serving weights ($w_{\text{rev}}, w_{\text{eng}}$). MSE diffs are relative to the revenue-only baseline (MTSH $\alpha=0.99$); recall diffs are relative to the engagement-only baseline (MTSH $\alpha=0.01$).

Model	MSE diff (%)	in-batch-2048 @ 10 diff (%)	offline-KNN @ 6400 diff (%)
MTSH ($\alpha=0.01$)	+6905.80	0.00	0.00
MTSH ($\alpha=0.99$)	0.00	-28.24	-21.19
JRC [25]	+5307.25	-0.03	-0.45
ESP (0.5, 0.5)	+52.17	-1.10	-1.29
ESP (0.8, 0.2)	+42.03	-0.67	-1.21
ESP (0.9, 0.1)	+23.19	-0.84	-0.76
ESP (0.99, 0.01)	+4.35	-0.54	-0.17

we validate that the partitioning principle remains robust across different architectures and delivers business impact at scale.

6.1 Objectives and Labels

In our production system, we jointly optimize 3 objectives.

Engagement. We utilize the user interaction data (135M clicks, 12M applies, 3M saves over 30 days) with in-batch negative sampling [9] and random easy negatives [32]. The loss is InfoNCE with temperature $T = 0.01$.

Revenue. We estimate revenue as $P(\text{click}|\text{impression}) \times \text{bid}$ under CPC pricing. We leverage knowledge distillation [7] from a high-precision L2 ranker to generate soft pClick labels via counterfactual data [4]. Training uses MSE loss against these pseudo-labels.

Fitness. We utilize soft labels generated from an LLM fine-tuned on 25K human annotations, quantifying how well a job matches a seeker’s qualifications. Binary labels are derived by thresholding, and training uses BCE loss.

6.2 Offline Results: ESP vs. MTSH

Table 5 compares architectures for jointly optimizing engagement and revenue. MTSH uses a single embedding with a weighted training loss (α : revenue loss weight). JRC [25] is a single-vector baseline that learns dual member embeddings for joint recall and calibration; like MTSH it cannot retune the engagement/revenue trade-off at serving time. ESP uses partitioned subspaces with per-objective serving-time weights ($w_{\text{rev}}, w_{\text{eng}}$).

MTSH shows severe interference at scale: emphasizing revenue ($\alpha=0.99$) loses 28.24% in-batch recall and 21.19% offline KNN recall. *At matched revenue emphasis* (MTSH $\alpha=0.99$ vs. ESP $w_{\text{rev}}=0.99$), MTSH loses 28.24% in-batch recall while ESP loses only 0.54%, a $\approx 50\times$ improvement that quantifies the interference that partitioning eliminates at production scale. JRC mitigates the recall loss (-0.03%) but inherits MTSH’s lack of serving-time tunability and pays a large MSE penalty (+5307.25%).

Scaling to $k=3$ in production. We add a fitness subspace alongside engagement and revenue. The fitness subspace achieves AUC 0.938

Table 6: Online A/B Test (d_1 : revenue, d_2 : engagement)

Weight w_{d_1}	Revenue	Job Application	Click Rate
0.1	+0.91%	+3.30%	+0.04%
0.2	+2.31%	+3.20%	+0.37%
0.3	+4.01%	+3.95%	+0.92%
0.5	+7.36%	+1.88%	+0.61%

on held-out fitness annotations, while engagement and revenue metrics remain within parity of the $k=2$ ESP configuration above. This validates that ESP’s per-subspace partitioning scales to three high-stakes objectives without measurable cross-objective regression, and provides the production counterpart to the open-source $k=3$ proof-of-concept in Section 5.7.

6.3 Online A/B Test

We deployed the ESP model trained with $\alpha = 0.99$ and varied the serving-time weight w_{d_1} for revenue, with $w_{d_2} = 1 - w_{d_1}$ controlling engagement. Table 6 reports statistically significant results ($p < 0.05$) from a 7-day online A/B test.

At $w_{d_1} = 0.3$ (the chosen production operating point), the system achieves simultaneous gains across all metrics; post-ramp analysis at this configuration shows an 8.95% reduction in the no-fit rate — the rate at which recommended jobs are flagged as poorly matched to seeker qualifications by downstream filters and human review. *A single trained ESP model spans four production operating points without retraining*, with revenue gains ranging from +0.91% to +7.36% and complementary engagement effects. The serving-time knob translates directly into deployable business levers, allowing product teams to rebalance objectives in response to evolving priorities without model retraining or infrastructure changes.

7 Discussion and Future Work

Serving Latency (Transformer-ESP). For the Transformer-ESP variant evaluated on the open-source benchmark (0.6B encoder on H200 GPUs), ESP introduces zero latency overhead: the causal-only variant (flash attention [5]) adds $<0.1\%$ over standard encoding, and the masked variant (SDPA) [23, 24] is 12% *faster* than flash attention on short sequences (22.4ms vs. 19.6ms per document). The custom 4D mask adds negligible cost beyond the SDPA switch itself. The production DNN-ESP system is much smaller and uses a different inference stack; we report production retrieval-quality results in Section 6 and refer readers to [12] for system-level latency.

Attribute Tagging as a Separate Problem. ESP decomposes multi-objective retrieval into two sub-problems: (1) entity/attribute tagging (upstream), and (2) partitioned embedding matching (ESP). The tagging model extracts the same typed attribute on both query and document sides. This is standard practice in production: job search systems extract job titles, e-commerce systems extract brand names. It is not a limitation of ESP but rather a natural pipeline decomposition. For queries without tagged attributes, the attribute segment is omitted and ESP gracefully reduces to standard single-embedding retrieval.

Infrastructure Prerequisite: Exhaustive KNN. ESP’s weighted dot product $\sum_i w_i \langle \mathbf{e}_i, \mathbf{e}'_i \rangle$ is incompatible with partition-based ANN indices such as IVFPQ [11], whose clustering and residual quantization presuppose a single coherent embedding geometry. A concatenation of independently partitioned subspaces does not satisfy this assumption. Recent advances in GPU accelerated exhaustive retrieval [12, 21] make ESP practical at web scale by enabling exact serving time reweighting over a unified index. As a result, ESP should be viewed as a joint model and systems design enabled by modern computation hardware.

Future Work. ESP opens several promising directions for future research in multi-objective retrieval. First, an important next step is evaluating settings where multiple objectives simultaneously require high-resolution embeddings, such as balancing semantic precision against topical diversity. Developing benchmarks with such high-entropy objective conflicts would provide a stronger test of the representational limits estimated by our theory. Second, we plan to improve the calibration of per-subspace scores using techniques such as post-hoc scaling or isotonic regression [3], which could simplify control over Pareto-optimal operating points as the number of objectives k grows. Extending ESP to larger numbers of heterogeneous objectives also remains an important practical direction. Finally, several theoretical and systems challenges remain open, including tightening the dimension allocation bounds in Theorem 5, improving cross-objective gradient flow in the shared encoder trunk, and developing more efficient segment-aware attention masking compatible with FlashAttention-style [5] kernels. We also plan to continue improving the benchmark infrastructure introduced in this work to support more rigorous and reproducible research on multi-objective retrieval systems.

8 Conclusion

We formalized the geometric limitation of single-embedding multi-objective bi-encoder retrieval and proposed ESP. The core practical insight is the *no-knob* problem: a single embedding provides no mechanism to control the trade-off between competing objectives at serving time, and the interference ceiling is architectural, not a training artifact. ESP resolves this by decomposing the embedding into isolated subspaces with a tunable weighted dot product. Unlike multi-head retrieval architectures that require per-objective ANN indices and a post-hoc merge, ESP serves via a single concatenated index and a single weighted dot product, enabled by GPU-accelerated exhaustive KNN [12]. In production at LinkedIn’s job matching platform (70M+ weekly active users), ESP yields +4.0% revenue, +3.95% job applications, and −8.95% no-fit rate from a single trained model spanning multiple operating points without retraining.

Acknowledgments

We thank the many talented engineers across LinkedIn whose dedication made this work possible, and our Product and Engineering leadership for their support.

References

- [1] Prabhat Agarwal, Minhazul Islam Sk, Nikil Pancha, Kurchi Subhra Hazra, Jiajing Xu, and Chuck Rosenberg. 2024. OmniSearchSage: Multi-Task Multi-Entity Embeddings for Pinterest Search. In *Companion Proceedings of the ACM Web Conference 2024*. Association for Computing Machinery, New York, NY, USA, 121–130. doi:10.1145/3589335.3648309
- [2] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. MS MARCO: A Human-Generated Machine Reading Comprehension Dataset. arXiv:1611.09268 [cs.CL] doi:10.48550/arXiv.1611.09268
- [3] Richard E. Barlow and Hugh D. Brunk. 1972. The Isotonic Regression Problem and Its Dual. *J. Amer. Statist. Assoc.* 67, 337 (1972), 140–147. doi:10.1080/01621459.1972.10481216
- [4] Léon Bottou, Jonas Peters, Joaquin Quiñero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. 2013. Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising. *Journal of Machine Learning Research* 14 (2013), 3207–3260.
- [5] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In *Advances in Neural Information Processing Systems*, Vol. 35. Curran Associates, Inc., New Orleans, LA, USA, 16344–16359. doi:10.52202/068431-1189
- [6] Carl Eckart and Gale Young. 1936. The Approximation of One Matrix by Another of Lower Rank. *Psychometrika* 1, 3 (1936), 211–218. doi:10.1007/BF02288367
- [7] Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. 2021. Knowledge Distillation: A Survey. *International Journal of Computer Vision* 129, 6 (2021), 1789–1819. doi:10.1007/s11263-021-01453-z
- [8] Zhuoning Guo, Guangxing Chen, Qian Gao, Xiaochao Liao, Jianjia Zheng, Lu Shen, and Hao Liu. 2026. GPU-Accelerated Multi-Relational Parallel Graph Retrieval for Web-Scale Recommendations. In *Proceedings of the ACM Web Conference 2026*. Association for Computing Machinery, New York, NY, USA, 7769–7778. doi:10.1145/3774904.3792788
- [9] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-Based Recommendations with Recurrent Neural Networks. In *4th International Conference on Learning Representations*. San Juan, Puerto Rico. arXiv:1511.06939
- [10] Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. 2020. Embedding-Based Retrieval in Facebook Search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, New York, NY, USA, 2553–2561. doi:10.1145/3394486.3403305
- [11] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. 2011. Product Quantization for Nearest Neighbor Search. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 1 (2011), 117–128. doi:10.1109/TPAMI.2010.57
- [12] Yuchin Juan, Jianqiang Shen, Shaobo Zhang, Qianqi Shen, Caleb Johnson, Luke Simon, Liangjie Hong, and Wenjing Zhang. 2025. Scaling Retrieval for Web-Scale Recommenders: Lessons from Inverted Indexes to Embedding Search. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*. Association for Computing Machinery, New York, NY, USA, 1066–1069. doi:10.1145/3705328.3748116
- [13] Vladimir Karpukhin, Barlas Ögüz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online, 6769–6781. doi:10.18653/v1/2020.emnlp-main.550
- [14] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 39–48. doi:10.1145/3397271.3401075
- [15] Weize Kong, Swaraj Khadanga, Cheng Li, Shaleen Kumar Gupta, Mingyang Zhang, Wensong Xu, and Michael Bendersky. 2022. Multi-Aspect Dense Retrieval. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, NY, USA, 3178–3186. doi:10.1145/3534678.3539137
- [16] Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. 2022. Matryoshka Representation Learning. In *Advances in Neural Information Processing Systems*, Vol. 35. Curran Associates, Inc., New Orleans, LA, USA, 30233–30249. doi:10.52202/068431-2192
- [17] Sen Li, Fuyun Lv, Taiwei Jin, Guli Lin, Keping Yang, Xiaoyi Zeng, Xiao-Ming Wu, and Qianli Ma. 2021. Embedding-Based Product Retrieval in Taobao Search. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, New York, NY, USA, 3181–3189. doi:10.1145/3447548.3467101
- [18] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards General Text Embeddings with Multi-Stage Contrastive Learning. arXiv:2308.03281 [cs.CL] doi:10.48550/arXiv.2308.03281

- [19] Xingxian Liu, Dongshuai Li, Jiahui Wan, Tao Wen, Gui Ling, Yuliang Yan, Fuyu Lv, Dan Ou, Haihong Tang, and Bo Zheng. 2025. Retrieval-GRPO: A Multi-Objective Reinforcement Learning Framework for Dense Retrieval in Taobao Search. arXiv:2511.13885 [cs.IR] doi:10.48550/arXiv.2511.13885
- [20] Jean Maillard, Vladimir Karpukhin, Fabio Petroni, Wen-tau Yih, Barlas Öğuz, Veselin Stoyanov, and Gargi Ghosh. 2021. Multi-Task Retrieval for Knowledge-Intensive Tasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 1098–1111. doi:10.18653/v1/2021.acl-long.89
- [21] Yu A. Malkov and Dmitry A. Yashunin. 2020. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 4 (2020), 824–836. doi:10.1109/TPAMI.2018.2889473
- [22] Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B. Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-T5 (ST5): Scalable Sentence Encoders from Pre-Trained Text-to-Text Models. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, Dublin, Ireland, 1864–1874. doi:10.18653/v1/2022.findings-acl.146
- [23] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. 2021. A Review on the Attention Mechanism of Deep Learning. *Neurocomputing* 452 (2021), 48–62. doi:10.1016/j.neucom.2021.03.091
- [24] Terence D. Sanger. 2026. Scaled Dot-Product Attention Implements Projection of Inputs onto a Common Surface. arXiv:2602.02521 [cs.LG] doi:10.48550/arXiv.2602.02521
- [25] Xiang-Rong Sheng, Jingyue Gao, Yueyao Cheng, Siran Yang, Shuguang Han, Hongbo Deng, Yuning Jiang, Jian Xu, and Bo Zheng. 2023. Joint Optimization of Ranking and Calibration with Contextualized Hybrid Model. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, NY, USA, 4813–4822. doi:10.1145/3580305.3599851
- [26] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. RoFormer: Enhanced Transformer with Rotary Position Embedding. *Neurocomputing* 568 (2024), 127063. doi:10.1016/j.neucom.2023.127063
- [27] Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 296–310. doi:10.18653/v1/2021.naacl-main.28
- [28] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text Embeddings by Weakly-Supervised Contrastive Pre-Training. arXiv:2212.03533 [cs.CL] doi:10.48550/arXiv.2212.03533
- [29] Per-Åke Wedin. 1972. Perturbation Bounds in Connection with Singular Value Decomposition. *BIT Numerical Mathematics* 12, 1 (1972), 99–111. doi:10.1007/BF01932678
- [30] Orion Weller, Michael Boratko, Iftekhhar Naim, and Jinhyuk Lee. 2026. On the Theoretical Limitations of Embedding-Based Retrieval. In *International Conference on Learning Representations*. arXiv:2508.21038 <https://openreview.net/forum?id=k9CzlvzfaA>
- [31] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *International Conference on Learning Representations*. arXiv:2007.00808 <https://openreview.net/forum?id=zeFrfgYzLn>
- [32] Ji Yang, Xinyang Yi, Derek Zhiyuan Cheng, Lichan Hong, Yang Li, Simon Xiaoming Wang, Taibai Xu, and Ed H. Chi. 2020. Mixed Negative Sampling for Learning Two-Tower Neural Networks in Recommendations. In *Companion Proceedings of the Web Conference 2020*. Association for Computing Machinery, New York, NY, USA, 441–447. doi:10.1145/3366424.3386195
- [33] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient Surgery for Multi-Task Learning. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., Online, 5824–5836. arXiv:2001.06782
- [34] 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*. Association for Computing Machinery, New York, NY, USA, 5194–5204. doi:10.1145/3711896.3737255
- [35] Jianjin Zhang, Zheng Liu, Weihao Han, Shitao Xiao, Ruicheng Zheng, Yingxia Shao, Hao Sun, Hanqing Zhu, Premkumar Srinivasan, Weiwei Deng, Qi Zhang, and Xing Xie. 2022. Uni-Retriever: Towards Learning the Unified Embedding-Based Retriever in Bing Sponsored Search. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, NY, USA, 4493–4501. doi:10.1145/3534678.3539212
- [36] Peitian Zhang, Zheng Liu, Shitao Xiao, Zhicheng Dou, and Jian-Yun Nie. 2024. A Multi-Task Embedder for Retrieval-Augmented LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 3537–3553. doi:10.18653/v1/2024.acl-long.194
- [37] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. arXiv:2506.05176 [cs.CL] doi:10.48550/arXiv.2506.05176