INSPIRE: Intent-aware Neural Sponsored Product Retrieval for E-commerce

Shasvat Desai[1,*], Hong Yao[1], Utkarsh Porwal[1] and Kuang-chih Lee[1]

[1]*Walmart Global Tech, USA*

Abstract

Walmart holds the largest share of the U.S. e-commerce grocery market, where food and beverage categories generate some of the highest search traffic and, consequently, drive a substantial portion of sponsored search revenue. At this scale, even small mismatches between user intent and retrieved products can lead to significant losses in both user engagement and monetization. Yet, understanding user intent in grocery search is inherently challenging. Queries are typically short, ambiguous, and highly diverse, often underspecifying critical preferences. For example, a query like schar white bread implicitly encodes a gluten-free preference through brand association, while queries such as chickpea pasta or oatmilk reflect underlying dietary preferences like gluten-free, plant-based, or lactose-free alternatives. Failing to capture these signals results in retrieving products that might be semantically similar but misaligned with the user's true needs.

From the advertiser's perspective, many products are explicitly designed to target specific intents—such as dietary preferences or size variants—and must be surfaced at the right moment to be effective. For example, a brand like Quest Nutrition, which sells high-protein, low-sugar snacks, wants its products to appear for queries like protein bars, low carb snacks, or keto snacks, even when these attributes might not be explicitly stated in the product title text. When retrieval systems fail to capture these intent signals, relevant products are not shown to the right users at the right time. From an advertiser's perspective, this means their products are missing high-intent opportunities where conversion is most likely. Over time, this leads to lower returns on ad spend, reduced trust in the platform, and potential advertiser attrition. Losing advertisers directly translates to a loss in advertising revenue and weakens the overall sponsored search ecosystem. This challenge is further amplified in sponsored search, where only a limited number of ad slots are available, making precise relevance essential.

Thus, we propose INSPIRE (Intent-aware Neural Sponsored Product Retrieval for E-commerce), an intent-aware retrieval framework for sponsored search that leverages structured intent signals to better align user queries with relevant food and beverage products. INSPIRE represents intent as a set of structured, multi-dimensional attributes derived from both user queries and product content, capturing explicit signals (e.g., brand, flavor) as well as implicit preferences (e.g., dietary constraints, cuisine types) that are often not directly expressed in queries. We develop a weakly supervised intent learning pipeline, where a large language model serves as a teacher to generate structured intent annotations from product titles and descriptions. We then distill these annotations by using them to finetune a lightweight student LLM model through LoRA based supervised finetuning (LoRA-SFT) that predicts intent attributes—such as brand, flavor, dietary preference, ingredient, product subtype, and cuisine type—at Walmart catalog scale. We then introduce an intent-augmented dense retrieval framework, where predicted intents are incorporated into query and product representations within a bi-encoder, enabling more precise matching between queries and sponsored products. To support real-world usage, we deploy the system as a scalable inference service. The distilled student model is served via a high-throughput API powered by vLLM, enabling efficient intent prediction over large product catalogs with low latency. This design ensures that intent-aware retrieval can be applied in production settings while maintaining efficiency and scalability.

Keywords

Intent aware retrieval, User behavior modeling, LoRA based LLM Supervised Finetuning for intent generation, Knowledge Distillation,

1. Introduction

Intent understanding for Walmart's grocery search is critical for two primary reasons: relevance and safety. Relevance is essential because Food and Beverages drives significant user traffic and revenue. Equally important is safety. Systems must respect dietary restrictions and ensure that retrieved products

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

*Corresponding author.

shasvat.desai@walmart.com (S. Desai); Hong.Yao0@walmart.com (H. Yao); utkarsh.porwal@gmail.com (U. Porwal); Kuangchih.Lee@walmart.com (K. Lee)

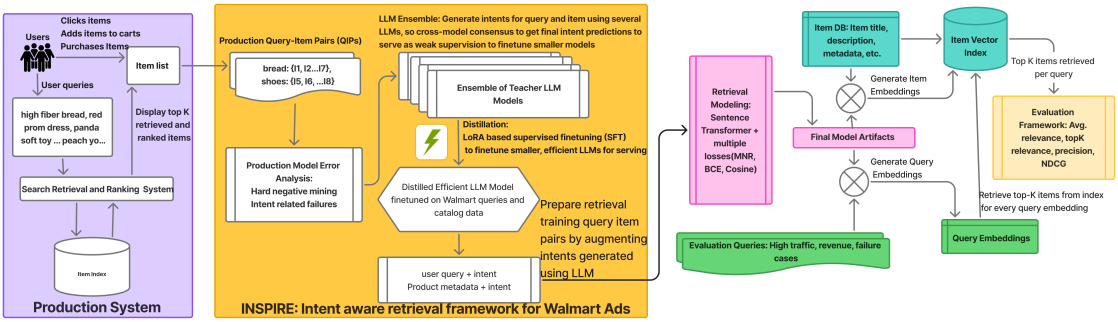

Figure 1: Overview of INSPIRE, an intent-aware retrieval framework for sponsored search. The system identifies failure cases in production query–item pairs, generates structured intents using an ensemble of teacher LLMs with cross-model consensus, and distills them into efficient student models via LoRA-based supervised fine-tuning. These intents are incorporated into query and item representations to train an intent-augmented dense retrieval model, enabling improved alignment between user queries and products.

are allergen-safe, as failures can lead to harmful user experiences. For example, a user with a peanut allergy searching for "nut free protein bars" should not be shown products containing peanuts, even if they are otherwise highly relevant. Intent understanding requires capturing both *explicit intents* (e.g., brand, flavor) and *implicit intents* (e.g., dietary preferences, cuisine type). When explicit signals are ignored—such as returning a different brand for "Coke vanilla" or mismatching flavor in "barbecue chips"—the results are immediately unsatisfactory and erode user trust.

The challenge is substantially greater for implicit intents, where users expect systems to infer constraints like sugar-free, dairy-free, or gluten-free even when they are only partially specified or entirely unstated. Crucially, this alignment must hold on both the *query side* and the *product side*: understanding what the user expresses is insufficient unless the system can also correctly interpret what the product represents. These signals are rarely explicit in product titles, requiring retrieval systems to reason over latent product knowledge rather than surface-level text. Brands often implicitly encode such properties—Schär is widely associated with certified gluten-free products, Nutpods with dairy-free and vegan offerings, and LILY'S with no-sugar-added chocolate—while ingredient-level cues further reinforce them (e.g., chickpea-based products are inherently gluten-free, Bomba rice is appropriate for paella whereas Arborio rice is not). Successfully retrieving the right items therefore demands a deeper alignment between implicit user constraints and equally implicit product attributes. As shown in Table 1, failing to capture these signals results in items that are lexically similar yet violate key constraints, making them practically unusable. Beyond preference, this also impacts safety and effort, as users are otherwise forced to inspect ingredient lists and nutritional information to verify suitability.

These challenges are further amplified in sponsored search. Ad engagement is inherently sparse—whether an ad is shown depends not only on relevance but also on campaign activity, auction dynamics, bid competitiveness, and advertiser budget. Moreover, the number of ad slots is significantly smaller than organic results, making precise relevance essential. Missing a relevant product therefore not only degrades user experience but also adversely impacts advertisers, leading to lost revenue, lower click-through rates, and reduced return on ad spend. Consequently, accurately modeling both query and product intents is critical to serving users effectively while sustaining a healthy advertiser ecosystem.

To address these challenges, we propose **INSPIRE (Intent-aware Neural Sponsored Product Retrieval for E-commerce)**, an intent-aware retrieval framework that leverages structured intent signals to better align queries with relevant food and beverage products. INSPIRE represents intent as a set of structured, multi-dimensional attributes derived from both queries and product content, capturing explicit signals (e.g., brand, flavor) as well as implicit preferences (e.g., dietary constraints, cuisine types).

We develop an INSPIRE with the following components: (i) **Failure identification**: perform hard negative mining and error analysis to surface query–item pairs where standard retrieval fails; (ii) **Intent ground truth annotation generation**: use multiple teacher LLMs to generate candidate intent ground truth labels from query and product content for weak supervision of student LLM ; (iii) **Consensus building**: apply cross-model agreement to filter noise and derive a high-confidence set of intent labels; (iv) **Model distillation**: train a smaller, efficient student LLM [1, 2] via LoRA-based [3] supervised fine-tuning using high-confidence, consensus-based intent derived from multiple teacher models as weak supervision to generate intents for Walmart catalog; (v) **Intent aware embedding for retrieval**: incorporate these intents into the embedding generation step of the retrieval model; (vi) **Evaluation**: measure improvements in retrieval quality using metrics such as NDCG@K and Precision@K; (vii) **Deployment**: serve INSPIRE framework at scale through a vLLM-based [4] inference API.

Query	**User Intent**	**Relevant Item**	**Irrelevant Item**
paella rice	Short-grain rice, Spanish cuisine	Goya Bomba Rice product_subtype: short grain, cuisine_type: Spanish	Goya Arborio Rice product_subtype: short grain, cuisine_type: Italian
gluten free bread	Gluten-free	Schär Artisan Baker White Bread dietary_preference: gluten free	Pepperidge Farm White Bread dietary_preference: contains gluten
sugar free chocolate	Low-sugar / no added sugar	Lily's Dark Chocolate Bar dietary_preference: sugar free	Hershey's Milk Chocolate Bar dietary_preference: high sugar
no dairy creamer	Dairy-free	Nutpods Unsweetened French Vanilla Creamer dietary_preference: dairy free	Coffee Mate Original Creamer dietary_preference: contains dairy
gluten free protein pasta	Gluten-free, high protein	Banza Chickpea Pasta dietary_preference: gluten free, protein	Barilla Semolina Pasta dietary_preference: contains gluten

Table 1: Implicit intent in grocery search is often not captured by lexical or semantic similarity alone. For example, paella specifically requires Bomba rice due to its cooking properties, while Arborio is suited for risotto. Similarly, brand-level (Schär, Nutpods, Lily's) and ingredient-level (chickpea) knowledge encode dietary constraints such as gluten-free, dairy-free, and sugar-free. This makes it challenging to match the correct product to a query based on product names alone. Many items may appear semantically similar (red), but fail to satisfy the underlying intent, while fewer items truly align with the user's needs (green). For instance, both Bomba rice and Arborio rice may match the query "paella rice" at a surface level, but only Bomba rice correctly satisfies the intent (green), whereas Arborio represents a misleading but similar alternative (red). Modeling these latent intent signals enables systems to distinguish between such cases and correctly rank intent-aligned products higher, while ignoring them results in items that are similar in wording but practically unusable.

Our contributions are as follows:

1. We propose **INSPIRE**, an intent-aware retrieval framework that incorporates structured intent signals into dense retrieval to improve alignment between user queries and sponsored products. We predict the following intents as part of INSPIRE: brand, dietary preference, flavor, ingredient, cuisine type, product subtype, size value and, size unit.
2. We develop a **weakly supervised labeling pipeline** that leverages large teacher LLMs and cross-model consensus to generate high-quality, structured intent annotations.

3. We propose a **distillation framework** that transfers these intent signals into efficient student models via LoRA-based supervised fine-tuning, enabling scalable intent prediction at catalog scale.
4. We demonstrate through **empirical evaluation** that intent-aware retrieval significantly improves relevance, particularly for queries involving implicit intent which cannot be effectively retrieved by semantic match.
5. We have designed and are in the process of deploying a **scalable inference system** using vLLM for low-latency, high-throughput intent prediction at catalog scale.

2. Related Work

Intent understanding has moved beyond simple taxonomies toward representation learning methods that model deeper semantic and contextual signals. Early foundations defined core navigational, informational, and transactional intents [5], followed by work leveraging behavioral signals such as click-through and session data to capture contextual intent while addressing biases in interaction logs [6, 7, 8, 9]. In e-commerce, intent understanding has been studied through query modeling, term refinement, and behavioral prediction. Product-aware approaches align query representations with catalog signals, enabling the model to ground user intent in the space of available products rather than treating queries in isolation [10], while hierarchical modeling captures multi-level user needs, improving retrieval for underspecified queries [11]. Complementary work shows that identifying and refining intent-bearing query terms improves retrieval effectiveness by filtering noisy or non-informative tokens [12], and sequential models leverage temporal user interaction patterns to predict purchase intent, particularly in session-based settings [13]. More recent approaches incorporate task-specific pre-training objectives to jointly optimize intent detection and embedding-based retrieval, leading to more semantically aligned representations [14].

In parallel, advances in neural retrieval have shifted from interaction-based models such as DRMM and KNRM, which rely on explicit term-level matching signals [15, 16], to dense retrieval frameworks enabled by Transformer architectures [17] and Siamese encoders like Sentence-BERT [18]. These models map queries and documents into a shared latent space, allowing semantic matching beyond lexical overlap. Training strategies such as hard negative mining (ANCE) [19], large-scale dual-encoder training (DPR) [20], and denoising with cross-encoder supervision (RocketQA) [21] further improve retrieval quality, while late-interaction models like ColBERT preserve fine-grained token-level matching with efficient indexing [22]. More recently, large language models (LLMs) have been leveraged for deeper intent understanding and semantic alignment [23], enabling extraction of nuanced, implicit user preferences. These models are typically adapted for production through distillation [24] or parameter-efficient tuning methods such as LoRA [3], with alignment techniques like Self-Instruct improving their ability to capture human intent [25].

However, existing approaches largely treat query intent understanding, item representation, and retrieval as loosely coupled components or rely heavily on behavioral signals. In contrast, our work jointly models structured intents for both queries and items, explicitly capturing both explicit signals (e.g., attributes, constraints) and implicit signals (e.g., brand priors, use-cases) via a multi-LLM consensus framework. We further distill these signals into an efficient student model for scalable deployment and integrate them directly into dense retrieval representations, enabling fine-grained alignment between user intent and product representations.

3. Method

We present **INSPIRE**, an intent-aware retrieval framework that improves alignment between user queries and sponsored products by incorporating structured intent signals derived from both query and product content. Our approach consists of five key components detailed in this section.

3.1. Error Analysis of Query–Item Pairs from Production Logs

We begin by analyzing production query–item pairs (QIPs) from Walmart logs, including impressions, clicks, add-to-cart, and purchase signals. We perform hard negative mining and error analysis to identify failure cases where the current retrieval system surfaces items that are semantically similar but misaligned with user intent. These failure cases highlight scenarios where implicit or explicit intent signals are not adequately captured. We analyze query- and item-level intent distributions to understand the potential impact of intent-aware retrieval.

To systematically analyze retrieval failures, we first construct a labeled set of query–item pairs (QIPs) from production logs. Each QIP is annotated with an ordinal relevance score $Rel(q, i) \in \{0, 1, 2, 3, 4\}$, corresponding to 0 (Embarrassing), 1 (Bad), 2 (Okay), 3 (Good), and 4 (Excellent), reflecting the degree to which the item satisfies the user's intent. These labels are obtained using a cascade of relevance models (Gemma-1B, Gemma-2B, and LLaMA-3 8B) [26, 27], evaluated sequentially with confidence-based early exit. For uncertain hard cases where early exit is not possible, the predictions from all models are aggregated via majority voting, with ties resolved using the final-stage model. This cascade of relevance models is trained using 1.2M query-product pairs, collected over 18 months. It was evaluated on a held-out set of 100K pairs annotated by trained raters on a 5-point scale (0=Embarrassing through 4=Excellent and class distribution approximately balanced at 18–22% per class). The resulting relevance score is normalized to $[-1, 1]$ as:

$$rel\_score(q, i) = (Rel(q, i) - 2)/2. \tag{1}$$

Query-side analysis. Approximately 60% of queries in our evaluation set have an average relevance rating below 3, indicating that retrieved results are often only moderately relevant or poor. Among these underperforming queries, 70% contain identifiable intent signals, and ∼20% involve *implicit intents* that are not explicitly expressed in the query text. Furthermore, ∼55% of these queries belong to the Food category, where latent preferences such as dietary constraints and cuisine types are particularly prevalent. These findings suggest that a large fraction of retrieval failures stem from missing or misaligned intent signals, and that modeling query-side intent can significantly improve relevance.

Item-side analysis. On the product side, intent signals are highly pervasive. We observe that 98% of food items contain at least one intent attribute. Notably, 59% of items contain *implicit intents* that are not explicitly stated in the product title, but can be inferred from brand, ingredients, or metadata. This highlights that product representations often encode rich latent signals that are not captured by surface text alone.

Failure case analysis. Table 2 shows examples of low-performing queries with identifiable intent signals. Many such queries contain clear intent attributes (e.g., flavor, ingredient, brand, dietary preference), yet the retrieval system fails to surface relevant items. This is especially pronounced for implicit intents (e.g., ingredient-level or cuisine-level signals), where the absence of explicit modeling leads to poor alignment despite strong lexical overlap.

These observations reveal a systematic gap: both queries and items contain rich intent signals—many of which are implicit—yet current retrieval systems fail to leverage them effectively. This motivates the need for an intent-aware retrieval framework that explicitly models and aligns query and product intents.

3.2. Teacher LLM Intent Prediction and Consensus

For the identified QIPs, we use an ensemble of large language models (LLMs) to generate structured intent annotations from both queries and product metadata (e.g., title, description). Each model predicts intent attributes such as brand, flavor, dietary preference, product subtype, and cuisine type. To improve robustness and reduce noise, we apply cross-model consensus, retaining only high-confidence intent predictions that are agreed upon across multiple teacher models. We construct high-quality intent

annotations using a multi-stage pipeline that combines multiple LLMs, cross-model agreement, and validation.

Query	Avg. Rel.	Intent (key:value)	Explicit	Implicit
soymilk vanilla	1.086	flavor: vanilla, ingredient: soy	flavor	ingredient
raspberry vinaigrette	1.086	flavor: raspberry	flavor	–
hash brown frozen	1.086	storage: frozen, ingredient: potato	storage	ingredient
monterey jack	1.086	ingredient: cheese	–	ingredient
bissell crosswave clean solution	1.086	brand: bissell	brand	–
ice salt	1.086	ingredient: salt	–	ingredient
cream of mushroom soup 4 pack	1.088	ingredient: mushroom, package: 4 pack	package	ingredient
real good chicken nugget	1.088	ingredient: chicken	–	ingredient
fruit riot	1.088	brand: fruit riot	brand	–
hot pocket philly steak	1.088	flavor: philly steak	flavor	–
heineken light	1.09	brand: heineken	brand	–
vegetarian sausage	1.09	dietary_preference: vegetarian	dietary	–
naan bread	1.09	cuisine_type: indian	–	cuisine
lactose free cheese	1.09	dietary_preference: lactose free	dietary	–
vegan cream cheese	1.092	dietary_preference: vegan	dietary	–
pho	1.092	cuisine_type: vietnamese	–	cuisine
gluten free bagel	1.092	dietary_preference: gluten free	dietary	–
black olive	2.258	ingredient: olive, color: black	color	ingredient
chicken organic	2.258	ingredient: chicken, dietary_preference: organic	dietary	ingredient
mini m m candy	2.258	brand: m&m	brand	–
organic apple cider vinegar	2.258	dietary_preference: organic	dietary	–
tcl roku television 65	2.26	brand: tcl, size: 65	brand, size	–
buffalo pretzel	2.26	flavor: buffalo	flavor	–
fresh ginger	2.26	ingredient: ginger	–	ingredient

Table 2: Error analysis using log data: Examples of low-performing queries with extracted intent signals. Each query is associated with an *average relevance rating across the top 25 retrieved items*, where ratings lie on an ordinal scale from 0 (Embarrassing), 1 (Bad), 2 (Okay), 3 (Good), to 4 (Excellent). Many queries contain explicit and implicit intents, yet retrieval performance remains poor, highlighting the gap between intent presence and utilization.

Step 1: Multi-LLM Intent Prediction. For both queries and items, we obtain structured intent annotations independently from three LLMs: Gemma3 27B, LLaMA 3.1 8B, and Qwen3 8B. [28, 27, 26]Each model produces a structured output (e.g., JSON) capturing intent attributes such as brand, flavor, ingredient, size value, size unit, product subtype, dietary preference, and cuisine type, including both explicit and implicit signals. Table 3 shows show sample intent ground truth annotations

Step 2: Cross-Model Agreement and Consensus. To improve robustness, we compare intent annotations across models. We perform pairwise comparisons (Gemma–LLaMA, Gemma–Qwen, LLaMA–Qwen) [27, 26, 28]and compute: (i) token-level overlap for each intent field, and (ii) semantic similarity using embedding-based cosine similarity to capture lexically different but semantically equivalent predictions. We use predefined thresholds (0.5 for token overlap, 0.3 for embedding similarity) to determine agreement. Intent fields agreed upon by at least two models are retained as high-confidence signals, while fields with no agreement are discarded.

Query/Item	Brand	Dietary Preference	Flavor	Subtype	Size
Skinnygirl, Fat-Free, Sugar-Free Raspberry Vinaigrette Salad Dressing, 8 fl oz	skinnygirl	sugar free, vegan, fat free	raspberry	raspberry vinaigrette	8 fl oz
Artisana Organics, Cashew Cacao Spread, 8 oz (227 g)	artisana organics	vegan, kosher, gluten free	cacao	cashew cacao spread	8 oz (227 g)
Silk Vanilla Almond Milk	silk	lactose free, vegan	vanilla	almond milk	–
Mountain Dew Zero	mountain dew	sugar free	–	soda	–

Table 3
Structured attribute extraction focusing on key user-relevant intents such as brand, dietary preferences, flavor, subtype, and size.

Step 3: LLM-based Verification. We further validate the consensus intents using a verification step with a GPT 4.1. [29] Given the original query/item along with the consensus and conflicting annotations, the model is prompted to validate correctness, detect hallucinations, and suggest corrections or missing implicit intents. Based on this output, intents are automatically accepted, rejected, or flagged for further review.

Step 4: Manual Spot-Checks. We perform targeted human review on low-confidence or conflicting cases to ensure annotation quality. This step helps correct edge cases and calibrate the overall pipeline.

Data Normalization and Synonym Mapping. All inputs are standardized prior to intent extraction. We apply lowercasing, whitespace normalization, and lexical normalization techniques such as stemming (e.g., "running" → "run") and lemmatization (e.g., "better" → "good", "mice" → "mouse") to reduce variability and improve consistency across model predictions.

To further ensure consistency, we normalize units by mapping surface forms (e.g., "grams", "gms", "pounds", "floz") to canonical representations (e.g., *g*, *lb*, *oz*). This reduces sparsity and improves alignment between query and product text.

We also perform synonym normalization to capture semantically equivalent expressions that may not match lexically. Using dependency parsing, we identify modifier patterns and map them to canonical intent attributes (e.g., "unsalted" → *salt-free*, "unsweetened" or "no sugar" → *sugar-free*, "decaf" → *caffeine-free*).

Together, these normalization steps ensure that both explicit and implicit intent signals are represented consistently, improving the reliability of intent extraction and downstream retrieval alignment.

3.3. Distillation via Supervised Fine-Tuning using LoRA

We distill consensus-based intent annotations into a smaller, efficient student model (`microsoft/Phi-4-mini-instruct`) [30] via supervised fine-tuning (SFT) [31] with Low-Rank Adaptation (LoRA)[3], enabling parameter-efficient adaptation while retaining knowledge transferred from larger teacher models. This enables scalable intent prediction while retaining knowledge encoded by larger teacher models. The resulting model predicts structured intents for both queries and products at catalog scale.

SFT trains the model on input–output pairs by minimizing the negative log-likelihood (NLL) of the target sequence conditioned on the input. Each example consists of a prompt (system + user input) and a completion (structured intent output). Inputs are tokenized and concatenated prior to training.

Original Form	Canonical Form
unsalted	salt free
unsweetened	sugar free
no sugar	sugar free
decaf	caffeine free
decaffeinated	caffeine free
kg, kilogram, kilograms, kilo, kilos	kg
g, gram, grams, gm, gms, gr, grm	g
lb, lbs, pound, pounds	lb
oz, ounce, ounces, ozs, fl oz, floz, fluid ounce, fluid ounces, 0z, z	oz
l, lt, liter, liters, ltr, litre, litres	l
ml, milliliter, milliliters, millilitre, millilitres	ml
cup, cups	cup
pint, pints, pt, pts	pint
quart, quarts, qt	quart
gallon, gallons, gal	gallon
packet, packets, sachet, sachets	packet
can, cans, tin, canister	can

Table 4: Canonical normalization of surface forms. We map diverse lexical variations—including dietary descriptors (e.g., unsweetened → sugar free) and unit expressions (e.g., grams, gms → g)—to standardized representations. This reduces sparsity and enables consistent intent extraction across queries and product descriptions.

Training Objective. We use token-level cross-entropy loss with label shifting, where the model predicts the next token given previous tokens. Padding tokens are ignored via an ignore index (set to -100).

We fine-tune the student model (`microsoft/Phi-4-mini-instruct`) [30] using parameter-efficient Low-Rank Adaptation (LoRA) [3]. Training is performed for 3 epochs with an effective batch size of 16 (batch size 1 with gradient accumulation), using a learning rate of 4×10^{-4}. We use 1 H100 GPU to train this model.

We construct training examples as prompt–completion pairs, where the prompt consists of the formatted input (query or product text), and the completion corresponds to structured intent annotations. The input sequence is formed by concatenating the prompt and completion tokens.

Prompt Masking and Supervision. We apply prompt masking to ensure that the loss is computed only over the completion tokens. Specifically, tokens corresponding to the prompt are assigned an ignore index (-100), while completion tokens are supervised using standard cross-entropy loss. Padding tokens are also masked to avoid contributing to the loss.

Formally, given input sequence $x = [x_{\text{prompt}}, x_{\text{completion}}]$, the training objective minimizes:

$$\mathcal{L}_{\text{SFT}} = -\sum_{t \in \text{completion}} \log P(x_t \mid x_{<t}) \tag{2}$$

We tokenize prompt and completion separately and concatenate them to form a single sequence, truncated to a maximum length of 4096 tokens. We use gradient accumulation to simulate larger batch sizes under memory constraints and periodically checkpoint model weights. Padding tokens and prompt tokens are excluded from supervision to prevent degenerate learning behavior (e.g., copying prompts), ensuring that the model focuses on generating accurate intent outputs.

3.3.1. Intent Prediction Evaluation

We evaluate intent prediction at the attribute level using semantic matching between predicted and ground truth intents. For each intent field, we compute embeddings using a SentenceTransformer (MiniLM) model for both prediction and ground truth.

A prediction is considered a true positive if the cosine similarity exceeds a threshold of 0.6. Based on this, we define:

- **True Positive (TP):** $\text{sim}(\text{pred}, \text{GT}) > 0.6$
- **False Positive (FP):** prediction present but not aligned with ground truth
- **False Negative (FN):** ground truth present but prediction missing
- **True Negative (TN):** both prediction and ground truth absent

We compute precision, recall, and F1-score across all intent attributes.

Item Understanding Results The model achieves strong performance on product data, with overall precision of 0.95 and recall of 0.97.

Intent	Precision	Recall
cuisine type	0.86	0.91
dietary preference	0.92	0.93
flavor	0.90	0.91
brand	0.92	0.90
ingredient	0.91	0.98
product subtype	0.98	0.99
quantity	0.65	0.64
size value	0.98	0.99
size unit	0.97	0.98

Query Understanding Results For query intent prediction, the model achieves overall precision of 0.91 and recall of 0.93.

Intent	Precision	Recall
brand	0.98	0.97
dietary preference	0.94	0.91
flavor	0.81	0.91
ingredient	0.77	0.84
size value	0.92	0.89
size unit	0.93	0.94
product subtype	0.94	1.00
cuisine type	0.83	0.81

Next, we leverage the distilled student model to enable scalable intent-aware retrieval. While large teacher LLMs produce high-quality intent annotations, they are computationally expensive and unsuitable for large-scale inference over millions of query–item pairs. To address this, we distill their knowledge into a lightweight student model, which retains comparable intent prediction quality while enabling low-latency, high-throughput inference.

3.4. Retrieval Data Augmentation

We use the distilled student model to predict structured intents for both queries and products at scale. These predicted intents are then used to augment the retrieval training data.

Representation	Query	Relevant Item	Irrelevant Item
Original	gluten free bread	Schär Artisan Baker White Bread	Pepperidge Farm White Bread
Intent-Augmented	gluten free bread ; dietary preference: gluten free	Schär Artisan Baker White Bread ; dietary preference: gluten free, product subtype: bread	Pepperidge Farm White Bread ; dietary_preference: contains gluten, product_subtype: bread
Original	no dairy creamer	Nutpods French Vanilla Creamer	Coffee Mate Original Creamer
Intent-Augmented	no dairy creamer ; dietary preference: dairy free	Nutpods French Vanilla Creamer ; dietary preference: dairy free, flavor: vanilla	Coffee Mate Original Creamer ; dietary preference: contains dairy
Original	paella rice	Goya Bomba Rice	Goya Arborio Rice
Intent-Augmented	paella rice ; cuisine type: Spanish, product subtype: short grain rice	Goya Bomba Rice ; cuisine type: Spanish, product subtype: short grain rice	Goya Arborio Rice ; cuisine type: Italian, product subtype: short grain rice

Table 5: Intent-aware data augmentation. Adding structured intent attributes to both queries and products enables explicit alignment of user constraints with product characteristics, improving discrimination between relevant and lexically similar but unsuitable items. For example, for the query *gluten free bread*, it is not immediately clear which product is a better match just based on the title: *Schär Artisan Baker bread* or *Pepperidge bread*. However, once we augment the intent key values, it becomes more obvious - the relevant item shares the attribute *dietary preference: gluten free*, whereas the irrelevant item contains gluten despite similar surface form. This differentiaion enables higher lexical and semantic match between query and relevant item, enabling us to improve relevance of retrieved items

3.5. Retrieval Model Training

We train an intent-aware dense retrieval model that leverages structured intent signals to better align user queries with relevant products. Our approach consists of three key components: (i) constructing a unified supervision signal for query–item pairs that captures both semantic relevance and user preference, (ii) augmenting queries and items with predicted intent attributes to enable explicit intent-level matching, and (iii) training a bi-encoder retrieval model using these enriched representations and supervision signals. We describe each of these components in detail below.

QIP Supervision Signal. We train the retrieval model on query–item pairs (QIPs) (q, i) constructed from production logs, capturing real user interactions. Each QIP is assigned a continuous supervision score that reflects both semantic relevance and user preference. We train the retrieval model on query–item pairs (QIPs) (q, i), each assigned a continuous supervision score capturing both semantic relevance and user preference.

Each QIP is first assigned an ordinal relevance label $Rel(q, i) \in \{0, 1, 2, 3, 4\}$, representing the degree of intent fulfillment from *Embarrassing* to *Excellent*. These labels are obtained via a cascade of cross-encoder teacher models (Gemma-1B, Gemma-2B, LLaMA-3 8B) along with available human annotations. We normalize this label to:

$$rel\_score(q, i) = \frac{Rel(q, i) - 2}{2} \in [-1, 1] \tag{3}$$

We further incorporate user engagement signals derived from aggregated interactions (orders, add-to-cart, clicks, views). These signals are log-compressed, normalized per query, and smoothed to produce $\tilde{E}(q, i)$. Engagement is incorporated only for semantically relevant QIPs ($Rel(q, i) \geq 2$) to avoid promoting irrelevant but popular items.

The final supervision score is defined as:

$$y(q,i) = \text{clip}\left(\mu_{\text{rel}} \cdot rel\_score(q,i) + \lambda_{\text{eng}} \cdot \tilde{E}(q,i), 0, 1\right) \tag{4}$$

This formulation ensures that supervision reflects both intent-level semantic alignment and real user preferences.

As described in Section 3.4, to enable intent-aware retrieval, we augment both queries and items with structured intent attributes predicted by the distilled student model. Let x_q and x_i denote the original query and item text, and $\mathcal{I}(q)$ and $\mathcal{I}(i)$ denote their predicted intent sets. We construct:

$$\tilde{x}_q = x_q \parallel \mathcal{I}(q), \quad \tilde{x}_i = x_i \parallel \mathcal{I}(i) \tag{5}$$

This augmentation enables the model to explicitly capture alignment between user constraints (e.g., dietary preferences, cuisine type) and product attributes.

Training Data. Our training dataset consists of approximately 10M QIPs derived from production logs, combining relevance-labeled pairs and engagement-based signals. This large-scale dataset provides diverse supervision across both explicit and implicit intent scenarios.

Model and Training Objective. We adopt a bi-encoder architecture based on a MiniLM Sentence-Transformer. [18, 32, 33] Queries and items are encoded independently:

$$\mathbf{h}_q = f(\tilde{x}_q), \quad \mathbf{h}_i = f(\tilde{x}_i) \tag{6}$$

Relevance is computed via cosine similarity:

$$s(q,i) = \cos(\mathbf{h}_q, \mathbf{h}_i) \tag{7}$$

We train the model using a combination of Multiple Negatives Ranking (MNR) loss [19] and cosine regression loss:

$$\mathcal{L} = \mathcal{L}_{\text{MNR}} + \lambda \cdot \mathcal{L}_{\text{cosine}} \tag{8}$$

The MNR loss enforces separation between positive and in-batch negative pairs, while the cosine loss aligns predicted similarity scores with the continuous supervision target $y(q,i)$. This enables the model to distinguish between items that are lexically similar but differ in intent-level compatibility.

4. Evaluation

We evaluate the effectiveness of INSPIRE using standard retrieval metrics, including NDCG@K, Precision@K, and average relevance@K. Evaluation is performed on a curated set of high-traffic queries, revenue-driving queries, and known failure cases, demonstrating improvements in alignment, particularly for queries involving implicit intent. We are still in progress of deploying this system and plan to run an A/B test after the deployment.

4.1. Evaluation Protocol

We construct an evaluation set of 12000 queries that covers both head and tail traffic regimes. Queries are sampled from (i) high-traffic segment, (ii) high-revenue segments, (iii) long-tail queries, and (iv) historically under-performing queries. We build an item index over the full item inventory using FAISS [34] and retrieve the top-25 candidates for each query. We report metrics at K=25 because the sponsored search page has a fixed number of ad slots; in our setting, 25 is the optimal number of ad items for our application. Each retrieved query–item pair is annotated on a 5-point relevance scale. We use a hybrid judging setup consisting of (i) available human annotations and (ii) model-based annotations from a strong DeBERTa cross-encoder relevance model fine tuned on internal data for

Metric	Non-Intent	Intent-Aware	Relative Gain
Avg relevance@10	3.033	3.105	+2.38%
Median relevance@10	3.0	3.0	–
Avg relevance@25	3.01	3.08	+2.3%
Median relevance@25	3.0	3.0	–
Precision@K (Rel $\geq$ 3)			
Precision@1	0.812	0.846	+4.2%
Precision@10	0.767	0.798	+4.0%
Precision@25	0.741	0.768	+3.6%
NDCG@K			
NDCG@1	0.903	0.932	+3.2%
NDCG@10	0.872	0.895	+2.64%
NDCG@25	0.854	0.874	+2.3%
Relevance Distribution (%)			
Score = 4 (Excellent)	43.0	47.3	+10.0%
Score = 3 (Good)	32.6	33.7	+3.4%
Score = 2 (Okay)	8.4	5.5	-34.5%
Score = 1 (Bad)	13.0	12.0	-7.7%
Score = 0 (Embarrassing)	3.0	1.5	-50.0%

Table 6
Offline retrieval performance comparison between baseline (Non-Intent) and intent-aware retrieval. Intent-aware modeling improves ranking quality across all metrics, with the largest gains at top ranks and a clear shift toward higher-quality results.

pairs without human labels. Out of the 300,000 query-item pairs (QIPs) that we evaluate the model on, human annotations were obtained for 100K QIPs. We use the resulting 5-point labels to compute graded retrieval metrics. We report (i) average relevance score over the retrieved list, (ii) Precision@K, and (iii) NDCG@K

4.2. Quantitative Evaluation

Intent-aware retrieval consistently improves ranking quality across all evaluation metrics. As shown in Table 6, average relevance@10 increases from 3.033 to 3.105 (+2.38%), and average relevance@25 improves from 3.01 to 3.08 (+2.3%), indicating better overall alignment between retrieved items and user intent, both at top ranks and deeper in the list.

We also observe consistent gains in standard IR metrics. Precision@K improves across all cutoffs, with the largest gains at top ranks (Precision@1: +4.2%), indicating that intent-aware modeling is particularly effective at improving early precision. Similarly, NDCG@K improves across all ranks (NDCG@10: +2.64%, NDCG@25: +2.3%), demonstrating that relevant items are not only retrieved more often but are also ranked higher.

Importantly, intent modeling shifts the relevance distribution toward higher-quality results. The proportion of *Excellent* items (score 4) increases by +10.0%, while *Good* items (score 3) also see a modest increase (+3.4%). In contrast, mid- and low-quality results are significantly reduced: score 2 (Okay) decreases by -34.5%, score 1 (Bad) by -7.7%, and score 0 (Embarrassing) by -50.0%. This indicates that incorporating structured intents helps eliminate semantically similar but constraint-violating items

Overall, these results validate that intent-aware representations lead to more precise, better-ranked, and more user-aligned retrieval outcomes.

Setting	Query	Irrelevant Item	Score	Relevant Item	Score
Without Intent	*peanut free snack*	Planters Roasted Peanuts	0.57	MadeGood Granola Bars	0.38
With Intent	*peanut free snack*	Planters Roasted Peanuts ; dietary_preference: contains peanuts	0.49	MadeGood Granola Bars ; dietary_preference: peanut free	0.58
Without Intent	*gluten free pasta*	Barilla Farfalle Pasta	0.56	Banza Chickpea Pasta	0.37
With Intent	*gluten free pasta*	Barilla Farfalle Pasta ; dietary_preference: contains gluten	0.51	Banza Chickpea Pasta ; dietary_preference: gluten free	0.88
Without Intent	*soy free hamburger buns*	Nature's Own Brioche Buns	0.50	Canyon Bakehouse Buns	0.55
With Intent	*soy free hamburger buns*	Nature's Own Brioche Buns ; dietary_preference: contains soy	0.57	Canyon Bakehouse Buns ; dietary_preference: soy free	0.89
Without Intent	*sugar free chocolate*	Hershey's Milk Chocolate	0.61	Lily's Dark Chocolate	0.44
With Intent	*sugar free chocolate*	Hershey's Milk Chocolate ; dietary_preference: contains sugar	0.45	Lily's Dark Chocolate ; dietary_preference: sugar free	0.84
Without Intent	*no dairy creamer*	Coffee Mate Original	0.59	Nutpods Creamer	0.46
With Intent	*no dairy creamer*	Coffee Mate Original ; dietary_preference: dairy	0.51	Nutpods Creamer ; dietary_preference: dairy free	0.86

Table 7
Impact of intent augmentation on retrieval scoring. Each query is shown under two settings: without intent (top) and with intent (bottom). In the absence of intent signals, the model relies primarily on lexical and semantic similarity, often assigning higher scores to semantically similar but constraint-violating items and lower score to relevant items(red). After augmenting item representations with structured intent attributes (e.g., dietary_preference), the model prioritizes items that satisfy both explicit and implicit user constraints by scoring them higher than the irrelevant one. Relevant items that were previously under-ranked receive higher scores after intent augmentation (green), demonstrating improved alignment between user intent and retrieved products.

4.3. Qualitative Evaluation

Table 7 provides qualitative examples illustrating how intent-aware modeling corrects common retrieval failures.

Without intent modeling, the system often favors lexically or semantically similar items that violate user constraints. For instance, for the query *"peanut free snack"*, a peanut-based product is incorrectly scored higher due to lexical overlap. Similarly, for *"gluten free pasta"*, traditional wheat pasta is preferred over a gluten-free alternative.

With intent augmentation, the model explicitly captures structured constraints such as *dietary_preference* (e.g., peanut-free, gluten-free, soy-free). This enables correct ranking of items that satisfy user requirements, such as allergen-free snacks, and chickpea-based pasta

These examples highlight that intent-aware retrieval enables constraint-level reasoning beyond surface similarity, leading to more accurate and user-aligned results.

Together, the quantitative and qualitative results demonstrate that modeling structured intents

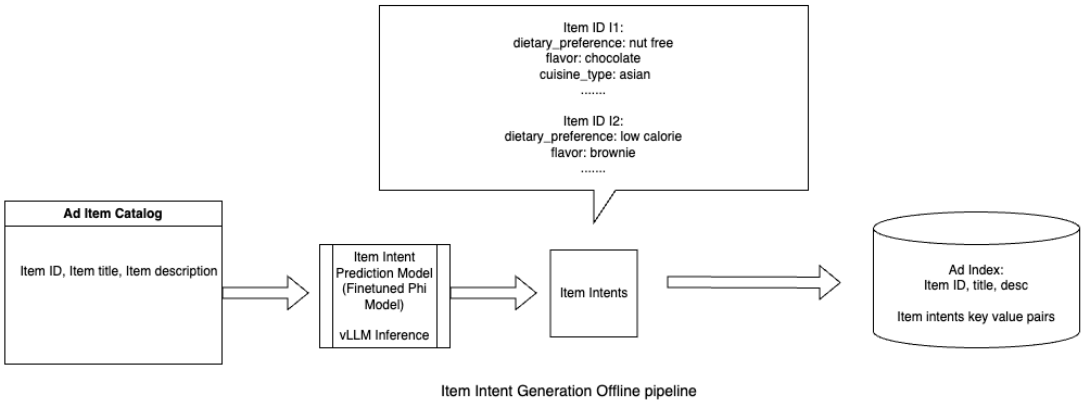

Figure 2: Offline pipelines for generating structured intents for items and queries. Item intents are extracted from catalog metadata and stored in the index. Query Intents come from an upstream intent prediction model(not shown here).

bridges the gap between intent presence and intent utilization in retrieval systems.

5. Serving Architecture

We are in process of developing and deploying this system to production and plan to run A/B test after that. At serving time, our system combines pre-computed item and query intents to enable intent-aware retrieval.

5.1. Offline Intent Generation Pipelines

As shown in Figure 2, we run a large-scale vLLM-based inference pipeline over the entire catalog to generate structured *item intents* (e.g., dietary preference, flavor, cuisine). We plan to store these intents in an *Item Intent Store* and inject them into the item index as additional structured features. On the query side, we plan to leverage intents predicted by another team to avoid duplication of effort.

Item intent generation. Each item (title and description) from the catalog is processed using a fine-tuned LLM to extract structured attributes such as dietary preferences, flavor, and cuisine type. These intents are stored as key-value pairs and incorporated into the item index. To keep the index up-to-date, we run incremental refreshes on newly added or updated items every hour, and a full refresh over the entire catalog on a weekly basis. This step is performed offline to ensure scalability and avoid serving-time overhead.

5.2. Online Pipeline

As shown in Figure 3, when a user issues a query, we check the if the normalized query exists(using embedding lookup) in the Query Intent Cache. As part of phase-1, we plan to use exact match(ENN) with the possibility to use approximate nearest neighbor match(ANN) to expand the query converage. The query, along with its inferred intents, is passed into the retrieval system, which leverages both semantic signals and intent alignment.

6. Limitations and Future Work

While our approach demonstrates consistent improvements in retrieval quality, it has several limitations. First, the paper only provides offline results. We plan to run A/B test after deploying this system to production and run an online test. Second, the quality of intent prediction is bounded by the teacher

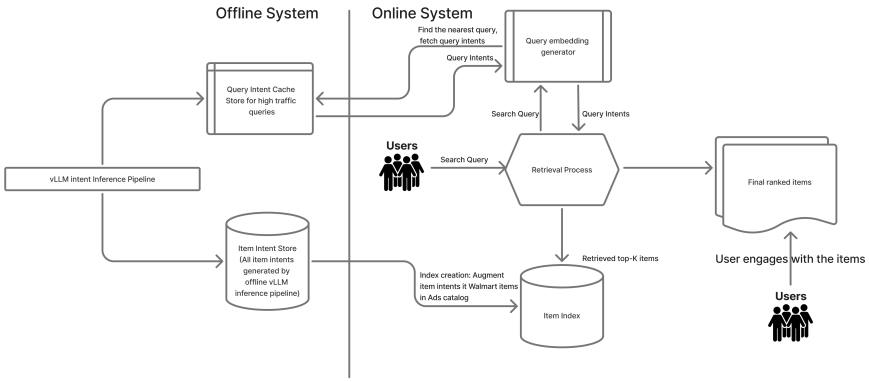

Figure 3: End-to-end serving architecture for intent-aware retrieval. The offline system generates item intents and caches query intents for frequent queries, while the online system performs real-time query intent inference and retrieves candidates from an intent-augmented index to produce final ranked results.

models used in the distillation pipeline. Errors or biases in the LLM-generated annotations can propagate to the student model, particularly for ambiguous or long-tail queries. Third, our framework relies on a predefined schema of structured intents, which may not fully capture the richness and evolving nature of user needs, especially in open-ended or cross-domain queries. Fourth, although we distill the model for scalability, maintaining up-to-date intent annotations for rapidly changing catalogs introduces additional system overhead.

In terms of future directions, one promising avenue is to move beyond fixed schemas toward more flexible or dynamic intent representations that can adapt to new domains and emerging user behaviors. Incorporating user context—such as session signals or personalization—can further improve intent disambiguation, particularly for underspecified queries. Another direction is to explore tighter integration between intent modeling and ranking, for example through joint training objectives or end-to-end optimization. Finally, improving the robustness of intent extraction through better alignment techniques, multi-model consensus strategies, or human-in-the-loop validation could further enhance reliability and downstream performance.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT (GPT-5) and Grammarly for grammar and spelling checks. After using these tools/services, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] K. Acharya, A. Velasquez, H. H. Song, A survey on symbolic knowledge distillation of large language models, IEEE Transactions on Artificial Intelligence 5 (2024) 5928–5948.

[2] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531 (2015).

[3] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank adaptation of large language models., Iclr 1 (2022) 3.

[4] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, I. Stoica, Efficient memory management for large language model serving with pagedattention, in: Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles, 2023.

[5] A. Broder, A taxonomy of web search, in: ACM Sigir forum, volume 36, ACM New York, NY, USA, 2002, pp. 3–10.

[6] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, H. Li, Context-aware query suggestion by mining click-through and session data, in: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008, pp. 875–883.

[7] A. Chuklin, I. Markov, M. De Rijke, Click models for web search, Springer Nature, 2022.

[8] T. Joachims, A. Swaminathan, T. Schnabel, Unbiased learning-to-rank with biased feedback, in: Proceedings of the tenth ACM international conference on web search and data mining, 2017, pp. 781–789.

[9] S. Desai, M. O. F. Rokon, J. N. Acharya, I. Shah, H. Yao, U. Porwal, K.-c. Lee, Unified supervision for walmarts sponsored search retrieval via joint semantic relevance and behavioral engagement modeling, arXiv preprint arXiv:2604.07930 (2026).

[10] J. Zhao, H. Chen, D. Yin, A dynamic product-aware learning model for e-commerce query intent understanding, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019, pp. 1843–1852.

[11] A. Ahmadvand, S. Kallumadi, F. Javed, E. Agichtein, Jointmap: joint query intent understanding for modeling intent hierarchies in e-commerce search, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 1509–1512.

[12] S. Manchanda, M. Sharma, G. Karypis, Intent term selection and refinement in e-commerce queries, arXiv preprint arXiv:1908.08564 (2019).

[13] K. I. Diamantaras, M. Salampasis, A. Katsalis, K. Christantonis, Predicting shopping intent of e-commerce users using lstm recurrent neural networks., in: DATA, 2021, pp. 252–259.

[14] Y. Qiu, C. Zhao, H. Zhang, J. Zhuo, T. Li, X. Zhang, S. Wang, S. Xu, B. Long, W.-Y. Yang, Pre-training tasks for user intent detection and embedding retrieval in e-commerce search, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022, pp. 4424–4428.

[15] J. Guo, Y. Fan, Q. Ai, W. B. Croft, A deep relevance matching model for ad-hoc retrieval, in: Proceedings of the 25th ACM international on conference on information and knowledge management, 2016, pp. 55–64.

[16] C. Xiong, Z. Dai, J. Callan, Z. Liu, R. Power, End-to-end neural ad-hoc ranking with kernel pooling, in: Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval, 2017, pp. 55–64.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.

[18] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), 2019, pp. 3982–3992.

[19] L. Xiong, et al., Approximate nearest neighbor negative contrastive learning for dense text retrieval, in: ICLR, 2021.

[20] V. Karpukhin, B. Oguz, S. Min, et al., Dense passage retrieval for open-domain question answering, in: EMNLP, 2020.
[21] Y. Qu, et al., Rocketqa: An optimized training approach to dense passage retrieval, in: NAACL, 2021.
[22] O. Khattab, M. Zaharia, Colbert: Efficient and effective passage search via contextualized late interaction, in: SIGIR, 2020.
[23] X. Liu, W. Guo, H. Gao, B. Long, Deep search query intent understanding, arXiv preprint arXiv:2008.06759 (2020).
[24] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).
[25] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, H. Hajishirzi, Self-instruct: Aligning language models with self-generated instructions, in: Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers), 2023, pp. 13484–13508.
[26] G. DeepMind, Gemma: Open models based on gemini research, arXiv preprint arXiv:2403.08295 (2024).
[27] H. Touvron, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).
[28] J. Bai, et al., Qwen technical report, arXiv preprint arXiv:2309.16609 (2023).
[29] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).
[30] A. Abouelenin, A. Ashfaq, A. Atkinson, H. Awadalla, N. Bach, J. Bao, A. Benhaim, M. Cai, V. Chaudhary, C. Chen, et al., Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, arXiv preprint arXiv:2503.01743 (2025).
[31] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Advances in neural information processing systems 35 (2022) 27730–27744.
[32] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, Advances in neural information processing systems 33 (2020) 5776–5788.
[33] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Huggingface's transformers: State-of-the-art natural language processing, arXiv preprint arXiv:1910.03771 (2019).
[34] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with gpus, IEEE transactions on big data 7 (2019) 535–547.