# Progressive Reinforcement Framework for Reliable E-Commerce Search Relevance

**Anonymous ACL submission**

## Abstract

Query-product relevance prediction is a core task in e-commerce search, critically impacting user experience and platform efficiency. While early BERT-based models struggled with complex queries, recent LLM-based methods that combine Supervised Fine-Tuning (SFT) with Reinforcement Learning (RL) still face two fundamental limitations: label-only rewards yield correct answers via flawed reasoning, undermining generalization; rigid step-level templates enforce uniform reasoning across all queries, suffering from poor adaptability and inefficiency. These challenges are especially severe in high-stakes domains such as medical e-commerce, where incorrect reasoning can lead to harmful recommendations. In this work, we propose **PRSR** (**P**rogressive **R**einforcement Framework for Reliable E-Commerce **S**earch **R**elevance), which enhances the reliability and generalization of LLMs in complex and high-risk scenarios through two key innovations: (1) a knowledge-guided, difficulty-aware multi-agent labeling system that synthesizes high-quality training data at reduced cost; and (2) a two-stage reinforcement learning strategy combining label-consistency rewards with difficulty-aware implicit process rewards, followed by targeted refinement on high-risk instances. Experiments demonstrate that PRSR significantly outperforms baselines in both offline and online evaluations, delivering substantial business gains in real-world e-commerce applications.

## 1 Introduction

Query-product relevance prediction, which matches user intent to product semantics, is the core of large e-commerce search systems, directly impacting user satisfaction.

Previous query-product relevance methods (Zou et al., 2021; Nogueira and Cho, 2019; Yates et al., 2021), mainly adopted the BERT (Devlin et al., 2019)-based architecture. However, this architecture simplifies the complex alignment between user intent, product information, and domain-specific constraints (such as medical indications or contraindications in drug search) into a single similarity score, sacrificing the accurate alignment of fine-grained information and failing to model complex reasoning processes. For example, the query "gastric juice" may refer to either "insufficient secretion" (highly relevant to digestive aids) or "acid reflux" (irrelevant), yet BERT collapses such nuanced judgments into a single score without leveraging domain knowledge or explicit reasoning.

The emergence of large language models (LLMs) (Achiam et al., 2023) has reignited interest in modeling relevance through deep semantic understanding and logical reasoning. Trained on vast amounts of text, LLMs internalize extensive world knowledge, endowing them with remarkable capabilities in language comprehension and multi-step reasoning—enabling them to handle real-world search scenarios that are both semantically complex and highly domain-specific. Ma et al. (2024) fine-tuned LLMs to directly output discrete relevance labels. While such black-box approaches are straightforward to implement, they suffer from poor interpretability and limited generalization, especially on rare or complex queries. More recently, generative relevance models (GRMs) based on explicit reasoning have primarily relied on supervised fine-tuning over large-scale human-annotated datasets. However, this paradigm lacks a closed-loop capability for exploration, feedback, and iterative optimization. To go further, some studies have incorporated reinforcement learning with label-based rewards. However these methods often produce correct answers via flawed reasoning paths—i.e., the model arrives at the right conclusion through shortcuts rather than valid logic—undermining robustness and generalization across diverse scenarios. Most recently, stepwise reward schemes have

been proposed that enforce a uniform reasoning pattern across all types of queries. However, such rigid, one-size-fits-all strategies are not only inefficient but also poorly generalized, as they fail to account for the heterogeneous reasoning structures required by different queries.

However, in medical e-commerce, relevance errors carry genuine health risks—e.g., recommending aspirin for a child's fever can cause serious harm (Schrör, 2007)—making accurate reasoning paths and controllable supervision critical. Unlike lifestyle search, medical relevance is a safety decision that must respect clinical evidence, yet existing LLM methods remain prone to hallucinations and lack domain-specific reasoning guidance.

To address these challenges—flawed reasoning under sparse label-only rewards and rigid stepwise templates that generalize poorly—we introduce PRSR, which augments LLM inference with structured domain knowledge, supervises reasoning via difficulty-aware implicit process rewards, and refines high-risk behavior through staged RL.

Our main contributions are as follows:

- We propose a knowledge-guided reasoning reinforcement learning framework that replaces rigid stepwise templates with flexible implicit semantic process rewards, achieving stronger generalization across specialized e-commerce search scenarios.
- We propose a reliable automatic labeling pipeline that generates relevance labels, reasoning steps, and calibrated difficulty scores from multi-model consensus disagreement, by combining structured product knowledge and expert guidelines—reducing hallucination, annotation cost, and manual errors while staying well-adapted to the domain.
- We propose a progressive RL strategy that first applies difficulty-aware implicit process rewards to prevent shortcut reasoning on ambiguous cases, then performs targeted high-risk refinement to ensure safety-critical behavior without catastrophic forgetting.
- Finally, we conducted extensive offline evaluations and large-scale online A/B tests on a real medicine e-commerce platform, verifying the superiority of our method compared to the baseline.

## 2 Related Work

### 2.1 Traditional and Neural Approaches to Search Relevance

Early approaches to search relevance modeling primarily relied on *sparse retrieval* methods such as BM25 (Robertson et al., 2009), which estimate query–document similarity based on lexical term matching and statistical heuristics. While efficient and simple, they fundamentally lack semantic understanding and fail to handle synonymy, paraphrasing, or compositional queries—limitations that are especially pronounced in e-commerce settings where user intents are often ambiguous and product descriptions are highly heterogeneous.

To bridge this gap, neural ranking models emerged, leveraging deep learning to capture semantic correspondences. Pioneering work like the Deep Structured Semantic Model (DSSM) (Huang et al., 2013) learned distributed representations of queries and items from implicit feedback signals (e.g., clicks). Later, Transformer-based architectures—particularly BERT-style cross-encoders (Zou et al., 2021; Nogueira and Cho, 2019; Yates et al., 2021)—achieved state-of-the-art performance by jointly encoding query–item pairs and fine-tuning on relevance labels, enabling rich contextualized interaction modeling.

Despite their success on frequent ("head") queries, these dense neural rankers still face significant challenges in real-world e-commerce and local-service search: they generalize poorly to long-tail or compositional queries (e.g., "What medication should children take for a cold and fever"), struggle with out-of-domain items, and offer limited interpretability. These shortcomings have motivated a recent shift toward large language models (LLMs), which bring strong world knowledge, few-shot adaptability, and generative reasoning capabilities—enabling more robust, flexible, and explainable relevance judgment.

### 2.2 Reasoning Large Language Models

Reasoning LLMs (large language models) have become a hot research area in recent years. Their core objective is to improve the model's ability to perform logical reasoning, mathematical reasoning, multi-step reasoning, and causal inference in complex tasks. While traditional LLMs excel in language generation and question answering, they still have significant shortcomings in tasks requiring systematic, structured, or symbolic rea-

soning. Google first proposed to use Chain-of-Thought prompting (Wei et al., 2022) and fine-tuning by instruction dataset (Wei et al., 2021), which significantly improves the ability of large language models to perform complex reasoning. But these methods heavily rely on manually designed demonstrations; they are one-time feedforward generation and lack the closed-loop capability of exploration-feedback-optimization. The emergence of o1 (Jaech et al., 2024) and DeepSeek-R1 (Guo et al., 2025) sparked a boom in using critic-free reinforcement learning to enhance reasoning capabilities, which employs reinforcement learning algorithms such as PPO (Schulman et al., 2017), GRPO (Shao et al., 2024) to reward answers sampled from the model itself, thereby activating the model's inherent rich world knowledge and improving its performance on challenging reasoning problems in the fields of mathematics and programming. Subsequently, more reinforcement learning algorithms (Yu et al., 2025; Zheng et al., 2025) and reasoning models (Yang et al., 2025a; Team et al., 2025; Zeng et al., 2025a) were introduced, establishing the use of reinforcement learning algorithms to enhance the reasoning capabilities of large language models as a widely adopted technical paradigm.

### 2.3 Reasoning-Enhanced LLM for Search Relevance

Recently, an emerging line of work has leveraged Reasoning Large Language Models (LLMs) for e-commerce relevance search by decomposing end-to-end relevance scoring into interpretable intermediate reasoning steps. Industrial practice at Tencent (Ye et al., 2025) has demonstrated the feasibility of deploying LLM-based rerankers in large-scale search systems. By leveraging Chain-of-Thought (CoT)-style prompts, these rerankers improve robustness on long-tail queries, though they do not explicitly optimize the reasoning trajectory itself—failing to address the underlying logical consistency of the inference process.

These methods share inherent limitations of Reasoning LLMs in e-commerce relevance tasks: supervised fine-tuning (SFT) tends to mimic surface patterns in training data rather than acquiring generalizable reasoning capabilities, leading to poor performance on novel, reasoning-intensive long-tail queries. To mitigate these issues, researchers (Jiao et al., 2025; Lu et al., 2025; Dong et al., 2025; Zeng et al., 2025b) have turned to reinforcement learning (RL) frameworks, leveraging fine-grained supervised signals to explicitly optimize reasoning logic. For instance, TaoSR1 (Jiao et al., 2025), TaoSR-SHE (Jiao et al., 2025), and TaoSR-AGRL (Yang et al., 2025b) progressively extended CoT reasoning into interpretable relevance assessment, sequential examination, and adaptive RL guidance. More recently, Xiaohongshu's search team Zeng et al. (2025b) developed Stepwise Advantage Masking (SAM) to assign per-step advantage estimates for CoT reasoning in relevance ranking.

While these methods have advanced stepwise reward design for reasoning-aware relevance search, they still face practical constraints. In some business scenarios the reasoning process is too complex to be solved with a universal template and the stepwise template cannot reuse in other fields. Moreover, none of them explicitly inject structured domain knowledge (e.g., drug attributes or clinical guidelines) into the reasoning process, leaving the model prone to hallucinations on domain-specific queries.

## 3 Method

### 3.1 Overview

Our framework proceeds in three progressively harder stages: (1) **SFT with CoT** to instill basic domain knowledge and instruction following; (2) **GRPO with label and process rewards**, where difficulty weighting focuses learning on ambiguous cases; and (3) **targeted refinement on high-risk samples** to improve safety-critical behavior without catastrophic forgetting. This curriculum moves from easy to hard and from low-risk to high-risk, letting the model build reliable reasoning before tackling the hardest cases.

### 3.2 Task Definition

We model search relevance as an interpretable generative classification task. Given a user's search query and a candidate product, the model not only needs to determine whether the product is relevant, but also needs to use medical evidence to explain the reasons for the judgment. The model input combines the user query and candidate product with pharmacist-curated guidelines and structured drug attributes (indications, contraindications, age and prescription restrictions). The model needs to generate two parts of output: a three-tiered relevance label: highly relevant, partially relevant, or irrelevant, reflecting whether the drug is suitable for

user needs in clinical practice; a concise reason, combined with medical knowledge and business rules, to explain why this judgment is given. This design transforms the traditional implicit sorting problem into an explicit decision-making process, while balancing the accuracy and interpretability of the results.

### 3.3 Knowledge-Guided Difficulty-Aware Reliable Data Synthesis

We present a knowledge-guided framework for synthesizing high-quality, difficulty-aware training data for medical relevance modeling. Our approach integrates structured medicine knowledge with multi-model LLM annotations to produce not only accurate relevance labels but also calibrated difficulty scores that reflect the inherent ambiguity or complexity of each query–product pair.

**Knowledge-Enhanced Prompt Engineering.** In collaboration with licensed pharmacists, we formulate a coherent medical reasoning guidelines that unifies domain-specific judgment criteria—such as indications, contraindications, age restrictions, and prescription status—into a consistent decision framework. For instance, the guidelines dictate that a query–product pair should be labeled "highly relevant" if the user query specifies the generic name of a drug and the retrieved product shares the same generic name.

This prompt is encoded into a natural language prompt template $\Pi$ as shown in Figure 1, which dynamically fuses: (i) the user query $q \in \mathcal{Q}$ (e.g., "pain relief for pregnant women"), (ii) the product title $p \in \mathcal{P}$, (iii) subgraphs from a curated medicine knowledge graph $\mathcal{K}_p$, and (iv) the guidelines set $\mathcal{G}$ derived from pharmacist expertise.

Table 1 shows an example output. The content within the reasoning tags represents the CoT rationale, and the content within the <answer> tags represents the final relevance label.

Table 1: An example of the model output.

| **Output:** |
| --- |
| + The keyword "anti-inflammatory" is highly relevant to the product "[Sanjiu] Metronidazole Tablets 0.2g × 15 tablets × 2 blisters/box" because the drug's primary therapeutic effects—antibacterial, anti-inflammatory, and anti-infective—directly align with the user's search intent for "anti-inflammatory," satisfying the criteria for high relevance in the relevance assessment guidelines. |
| + |
| <answer> |
| Highly Relevant |
| </answer> |

**Multi-Model Annotation and Consensus Labeling.** We invoke multiple state-of-the-art LLMs (e.g., Qwen-Max, GPT-4) to annotate real-world query–product pairs using the prompt $\Pi(q, p, \mathcal{K}_p; \mathcal{G})$. Each model $\mathcal{M}_i$ outputs a tuple $y_i = (l_i, t_i)$, where $l_i \in \{\texttt{HR}, \texttt{PR}, \texttt{IR}\}$ denotes the predicted label (highly relevant, partially relevant, or irrelevant), and $t_i$ is the chain-of-thought rationale grounded in $\mathcal{K}_p$ and $\mathcal{G}$.

For cases where ground-truth relevance can be deterministically verified via exact matching against $\mathcal{K}_p$ (e.g., query matches product generic name), we fix the label to the correct value and request only rationale generation, ensuring factual alignment.

Crucially, rather than using raw model outputs directly, we construct a consensus-based synthetic dataset by combining confidence-weighted voting for labels with rationale generation from the most reliable model. Specifically, we first estimate each model's reliability by computing its accuracy on a human-annotated validation subset, yielding confidence scores $c_i \in [0, 1]$.

For each sample, the final relevance label $\hat{l}$ is determined by confidence-weighted voting, where $\mathcal{L} = \{\texttt{highly_relevant}, \texttt{partially_relevant}, \texttt{irrelevant}\}$ denotes the set of candidate labels:

$$\hat{l} = \arg\max_{k \in \mathcal{L}} \sum_{i: l_i = k} c_i.$$

The corresponding CoT rationale $\hat{t}$ is then generated by the highest-confidence model with voted label $\hat{l}$, using the prompt $\Pi'(q, p, \hat{l}, \mathcal{K}_p; \mathcal{G})$, where $\Pi'$ denotes a variant of $\Pi$ adapted to incorporate the input label $\hat{l}$. While this may appear to risk post-hoc rationalization, we argue it constitutes principled knowledge distillation: the consensus label is a high-quality voting signal, and the CoT prompt requires grounding in $\mathcal{K}_p$ and $\mathcal{G}$ so the rationale cites domain-specific evidence rather than fabricates—paralleling the paradigm in DeepSeek-R1 (Guo et al., 2025) and TaoSR1 (Dong et al., 2025).

Thus, $D_{\text{distill}} = \{(x, \hat{l}, \hat{t})\}$, where $x = \Pi(q, p, \mathcal{K}_p; \mathcal{G})$, serves as the foundation for all downstream training.

**Semantic Data Augmentation with Retrieve-augmented generation(RAG).** To improve coverage of edge cases, we augment $D_{\text{distill}}$ via approximate nearest neighbor (ANN)(Arya et al., 1998)

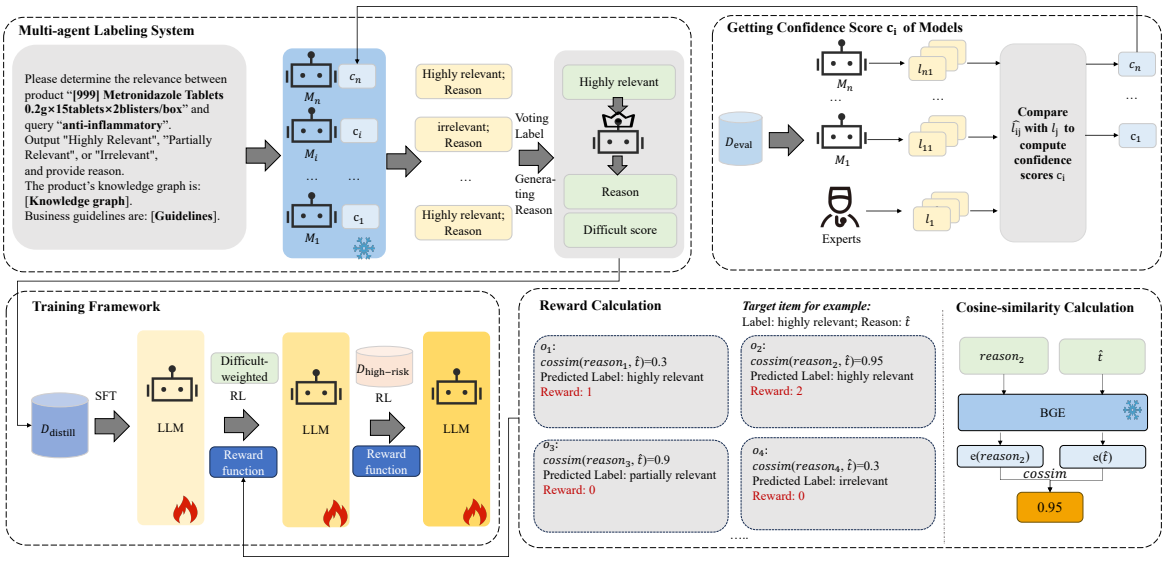


Figure 1: An overview of the proposed three-stage training framework, beginning with a multi-agent data synthesis pipeline, and detailing the reward function and model confidence scores.

retrieval over embeddings of logged clicked pairs. For each retrieved candidate $(q', p')$, we reuse the original label and rationale as guidance and prompt LLMs to generate new CoTs consistent with $\mathcal{K}_{p'}$ and $\mathcal{G}$. The augmented samples are incorporated into $D_{\text{distill}}$ only after passing consistency checks.

The final $D_{\text{distill}}$ thus consists of tuples $(q, p, \hat{l}, \hat{t})$, where the labels and rationales derive from the highest-confidence consensus. The degree of model disagreement during consensus naturally reflects sample difficulty; this signal is formalized as $d(x)$ in Section 4.3 to guide difficulty-aware RL weighting.

### 3.4 Supervised Fine-Tuning of Reasoning

We fine-tune a compact, deployable LLM (e.g., Qwen-8B or Llama-3-8B) on $D_{\text{distill}}$ to jointly predict the relevance label and generate the full Chain-of-Thought in an autoregressive manner. The training objective minimizes the negative log-likelihood:

$$\mathcal{L}_{\text{SFT}}(\theta) = -E_{(x,\hat{y})\sim D_{\text{distill}}}\left[\sum_{i=1}^{|\hat{y}|} \log \pi_\theta(\hat{y}_i \mid \hat{y}_{<i}, x)\right]. \tag{1}$$

This phase enables the model to internalize pharmacist-aligned reasoning patterns. However, empirical evaluation shows that SFT alone often leads to superficial mimicry—particularly on high-difficulty or safety-critical cases—motivating our subsequent reinforcement learning stage.

### 3.5 Safety-Aware Reinforcement Learning with Process-Guided Rewards

**Hybrid Reward for Label and Reasoning.** To address the limitations of SFT, we apply Group Relative Policy Optimization (GRPO) (Shao et al., 2024) with a reward function that jointly evaluates label correctness and reasoning fidelity. Unlike explicit stepwise matching that enforces rigid structural templates (Jiao et al., 2025), we employ an implicit semantic similarity reward via BGE embeddings, allowing flexible reasoning trajectories that are not penalized for valid but differently phrased logic.

Let $s_i = \cos\text{-sim}\big(\mathbf{e}(t_p^{(i)}), \mathbf{e}(\hat{t})\big)$. The reward $r_i$ is defined as:

$$r_i = \begin{cases} r_l + \beta r_p, & \text{if } l_p^{(i)} = \hat{l} \text{ and } s_i \geq \alpha, \\ r_l, & \text{if } l_p^{(i)} = \hat{l} \text{ and } s_i < \alpha, \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where $\mathbf{e}(t)$ denotes the sentence embedding of text $t$ obtained from the BGE model (Chen et al., 2024), which is a pre-trained and fixed embedding encoder optimized for semantic similarity tasks, and $\cos\text{-sim}(\mathbf{u}, \mathbf{v}) = \dfrac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\|\,\|\mathbf{v}\|}$ denotes the cosine similarity(Salton et al., 1975) between vectors $\mathbf{u}$ and $\mathbf{v}$.

The GRPO objective optimizes $\pi_\theta$ relative to a reference policy $\pi_{\text{ref}}$ (initialized as the SFT model):

$$\mathcal{J}_{\text{GRPO}}(\theta) = E_{x,o_i}\Big[w(x)\,\tfrac{1}{G}\sum_{i=1}^{G}\min\big(\rho_i\hat{A}_i, \text{cl}_\varepsilon(\rho_i)\hat{A}_i\big) - \gamma\,\text{KL}\big(\pi_\theta\|\pi_{\text{ref}}\big)\Big], \tag{3}$$

where $x \sim D_{\text{distill}}$, $o_i \sim \pi_{\theta_{\text{old}}}$, and $\text{cl}_\varepsilon(\rho) = \text{clip}(\rho, 1-\varepsilon, 1+\varepsilon)$.

where

$$\rho_i(\theta) = \frac{\pi_\theta(o_i \mid x)}{\pi_{\theta_{\text{old}}}(o_i \mid x)}. \tag{4}$$

with advantages normalized within each group:

$$\begin{aligned} \hat{A}_i &= \frac{r_i - \mu_r}{\sigma_r + \delta}, \\ \mu_r &= \frac{1}{G}\sum_{j=1}^{G} r_j, \\ \sigma_r &= \sqrt{\frac{1}{G}\sum_{j=1}^{G}(r_j - \mu_r)^2}. \end{aligned} \tag{5}$$

and $w(x)$ represents the advantage weighting for difficult samples:

$$w(x) = 1 + \omega d(x), \tag{6}$$

where

$$d(x) = \frac{\sum_{i=1}^{N} c_i \cdot I\big(l_i(x) \neq \hat{l}\big)}{\sum_{i=1}^{N} c_i}, \tag{7}$$

denotes a normalized disagreement score that reflects how much teacher models diverge from the consensus. The greater the divergence of the models, the higher this score becomes—indicating semantic ambiguity, rare medical limitations, or out-of-distribution situations.

**Why implicit semantic similarity?** We compare our implicit BGE-based process reward against two alternatives (Table 3). A dense variant that replaces the binary reward with a continuous similarity score yields notably lower F1, confirming that sparse, high-confidence signals provide more stable supervision. Second, an explicit step-wise matching baseline inspired by TaoSR (Jiao et al., 2025) extracts structured fields from the CoT and rewards character-level overlap. While this yields competitive 2-class F1, it underperforms on 3-class F1 (80.53% vs. 86.10%), suggesting that rigid structural matching penalizes valid but differently phrased reasoning.

**Targeted Refinement on High-Risk Scenarios.** In this section, we distinguish difficult examples—where the optimized model still shows low confidence or high output variance—and high-risk examples, where errors may cause serious issues in security, compliance, or user experience. While key business scenarios are often both difficult and high-risk due to their complexity, not all difficult examples carry significant risk (e.g., obscure queries). To align technical optimization with business priorities, we explicitly construct a high-risk sample set targeting critical scenarios and guide GRPO training toward them.

To further enhance robustness in safety-critical settings, we curate a specialized high-risk dataset $D_{\text{high_risk}}$, comprising historical failure cases and its most corresponding rule. Using its most corresponding rule to help model focus on the core problem, we generate high-fidelity labels and CoTs for these examples with the strategy described in Section 3.3.

We then perform an additional GRPO round on $D_{\text{high_risk}}$ mixed with $D_{\text{distill}}$, using the same hybrid reward function as the main GRPO stage and updating the reference model to the post-main-GRPO checkpoint. The GRPO objective is the same as Equation (3), where the advantage weighting is defined as:

$$w(x) = 1 + \omega_{\text{h}} I_{\text{hr}}(x) \cdot (\sigma_r(x)), \tag{8}$$

and:

- $I_{\text{hr}}(x) \in \{0, 1\}$ is the indicator function for high-risk samples in $D_{high_risk}$;
- $\omega_{\text{h}} \geq 0$ is the base weight for high-risk samples;
- $\sigma_r(x) = \text{std}\big(\{r(o_i)\}_{i=1}^{G}\big)$ is the reward variance across rollouts.

# 4 Experiments

## 4.1 Experimental Setup

### 4.1.1 Benchmarks and Metrics

For offline evaluation, our experiments are conducted on two test sets constructed from the training statistics shown in Appendix A: RANDOM and DIFFICULT. The RANDOM dataset is constructed by randomly sampling query–item pairs that incorporate all data synthesis strategies, ensuring broad coverage of typical cases. In contrast, the DIFFICULT dataset specifically includes pairs that were flagged as difficult (when $d$ is greater than a certain threshold). All labels in the test sets of both datasets were meticulously annotated

by professional pharmacists to ensure high-quality ground truth. If not mentioned in the result table, the default is to use the RANDOM dataset.

We use weighted-F1 to perform three-class and two-class evaluations on our model. Two-class classification categorizes partially relevant and irrelevant queries into one class, representing the baseline performance guarantee of the system (i.e., ensuring reliable filtering of irrelevant results), while three-class classification represents business objectives that can be finely distinguished.

All training hyperparameters and implementation details are provided in Appendix B; online evaluation metrics are defined in Appendix C.

#### 4.1.2 Dataset

We construct $D_{\text{distill}}$ (150k samples from search logs, augmentation, and rule-based matching) and $D_{\text{high_risk}}$ (20k failure cases) using the multi-agent labeling pipeline described in Section 3.3; detailed statistics are in Table 5.

#### 4.1.3 Models

We evaluate a series of baselines based on large language models (LLMs) of varying scales and prompt strategies. Since our proposed method involves multiple training stages, we also assess the model performance at each intermediate stage to analyze the contribution of individual components.

**Baselines.** We use teacher-model confidence-weighted voting as an upper-bound baseline. Two zero-shot variants of `Qwen3-8B-thinking` (Yang et al., 2025a) serve as source baselines: the vanilla model and a knowledge-augmented version that receives prompt $\Pi(q, p, \mathcal{K}_p; \mathcal{G})$ at inference time.

Four progressively stronger checkpoints trace the contribution of each component:

- **SFT-woCoT / SFT-CoT**: Fine-tuning on $(q, p, l)$ without / with chain-of-thought supervision.
- **RL-Label**: GRPO with label-based reward only ($\beta$=0).
- **RL-Process**: GRPO with our implicit BGE-based process reward added to the label reward.
- **RL-Process(CoT-pattern)**: A reference baseline implementing TaoSR-style stepwise reward (Jiao et al., 2025; Yang et al., 2025b): it extracts query intent, product entity, and grounding rule from the CoT and rewards character-level F1 overlap against teacher references.
- **RL-Risk(PRSR)**: A second GRPO stage on $D_{\text{high_risk}}$ mixed with the main training data, refining high-risk behavior without catastrophic forgetting.

**Note on External Baselines.** Direct comparison with TaoSR (Jiao et al., 2025; Yang et al., 2025b; Dong et al., 2025) or LORE (Lu et al., 2025) is infeasible: incompatible annotation schemes, proprietary datasets, and domain-specific reasoning templates prevent cross-domain transfer. We therefore reproduce their stepwise reward paradigm via **RL-Process(CoT-pattern)** on our own data, enabling a controlled comparison of explicit structural matching versus our implicit BGE-based semantic reward (Table 3).

### 4.2 Offline Results

Table 2 presents a comparison between PRSR and various baselines on both RANDOM and DIFFICULT test sets. Teacher model confidence scores are provided in Appendix D.

From the results, we draw the following observations:

- The design of prompts based on $\Pi(q, p, \mathcal{K}_p; \mathcal{G})$ is crucial, as it integrates domain knowledge $\mathcal{G}$ and product-specific context $\mathcal{K}_p$, guiding the model to produce results that are more in line with expectations based on business requirements and stable product knowledge. This gives the model initial domain capabilities and reduces hallucinations.
- Supervised fine-tuning based on Chain-of-Thought (SFT-CoT) can further improve the model's domain capabilities by learning from the teacher models' capabilities; however, it still falls short of the teacher model's upper limit.
- Introducing GRPO for label rewards guides the model to freely explore inference paths, effectively approximating the teacher model's capabilities; see Appendix D for teacher confidence/F1 scores.
- Introducing difficulty-weighted and reasoning process rewards to GRPO effectively guides training strategies, especially when dealing with complex or ambiguous samples, preventing the model from converging to incorrect inference paths. On the RANDOM set, this enables the model to exceed the teacher model's confidence/F1 score of 84.61% (Appendix D).
- By extending the two-stage GRPO framework with an adaptive weighting mechanism for high-risk samples, our method improves performance from critical boundary cases without sacrificing generalization to the original data distribution. A

Table 2: Performance comparison of main strategies.

| Strategy | Training Data | RANDOM | | DIFFICULT |
|---|---|---|---|---|
| | | F1 (3-class) | F1 (2-class) | F1 (3-class) |
| Qwen3-8B-source (baseline) | – | 56.13% | 85.30% | – |
| Qwen3-8B-with-knowledge | – | 74.80% | 86.50% | – |
| SFT-CoT | $D_{\text{distill}}$ | 76.70% | 82.00% | – |
| RL-label | $D_{\text{distill}}$ | 82.30% | 86.10% | – |
| RL-Process | $D_{\text{distill}}$ | 86.10% | 87.20% | 23.05% |
| RL-Risk(PRSR) | $D_{\text{distill}} + D_{\text{high_risk}}$ | 86.60% | 91.40% | 67.17% |

Table 3: Ablation study of training strategies.

| Strategy | Data | F1(3-cls) | F1(2-cls) |
|---|---|---|---|
| SFT-woCoT | $D_{distill}$ | 75.10% | 82.20% |
| SFT-CoT | $D_{distill}$ | 76.70% | 82.00% |
| RL-Label (base) | $D_{distill}$ | 76.20% | 87.20% |
| RL-Label (SFT init) | $D_{distill}$ | 82.30% | 86.10% |
| RL-Process (no diff.) | $D_{distill}$ | 81.20% | 87.10% |
| RL-Process (dense rwd) | $D_{distill}$ | 75.50% | 80.00% |
| RL-Process (CoT-pattern) | $D_{distill}$ | 80.53% | 88.51% |
| RL-Process | $D_{distill}$ | 86.10% | 87.20% |
| RL-Process | $D_{distill}$+$D_{hr}$ | 85.75% | 90.22% |
| RL-Risk | $D_{hr}$ | 78.60% | 85.80% |
| RL-Risk (PRSR) | $D_{distill}$+$D_{hr}$ | 86.60% | 91.40% |

significant gain is observed on the Difficult test set.

- Finally, RL-Risk (PRSR) outperforms the teacher model (84.61% F1; Appendix D) and all intermediate baselines (Table 2) on the RANDOM set, demonstrating that staged RL yields net improvements beyond simple teacher distillation.

### 4.3 Ablation Studies

Prompt engineering ablations are provided in Appendix E, and training-strategy ablations are summarized in Table 3.

Table 3 summarizes key training ablations. **(1)** SFT alone exhibits weak generalization, confirming the necessity of RL. **(2)** GRPO strongly benefits from SFT initialization. **(3)** Difficulty-aware process rewards outperform label-only and dense-reward variants; detailed reward-design comparisons are provided in Appendix F. **(4)** Two-stage training with mixed data prevents catastrophic forgetting while improving edge-case performance. **(5)** Naively mixing high-risk data into a single stage degrades generalization (86.10% $\rightarrow$ 85.75%), validating the staged design. Hyperparameter sensitivity is reported in Appendix G.

Table 4: Online performance comparison of main models.

| Model | PV ↑(%) | CTR ↑(%) |
|---|---|---|
| `Human (online)` | 0 | 0 |
| `RL-Risk(PRSR)` | +5.76 | +0.06 |

### 4.4 Online Deployment and A/B Test

PRSR was deployed on the A/B testing platform of a large medicine e-commerce search system. We randomly assigned 40% of online traffic to the test group (PRSR) and 60% to the baseline (human annotation), continuously monitoring business metrics over a minimum of seven days to mitigate traffic fluctuation. Compared with the human baseline, PRSR achieved a substantial PV increase while maintaining a close CTR level (Table 4). Given the large traffic volume, even these marginal differences are statistically detectable with high confidence, confirming the practical value of our approach.

## 5 Conclusion

We present PRSR, a reliable reinforcement learning framework for e-commerce search relevance. By combining knowledge-guided multi-agent data synthesis with a two-stage GRPO pipeline—label and difficulty-aware process rewards followed by high-risk refinement—our approach achieves strong generalization on challenging queries without catastrophic forgetting. Extensive offline and online evaluations confirm its practical value in real-world pharmaceutical search systems.

## Limitations

Despite the strong empirical results, PRSR remains fundamentally constrained by the quality of its teacher models. The multi-agent consensus pipeline aggregates reasoning from multiple LLMs to approximate expert-level labels, yet it implicitly bounds the student model's performance ceiling: if all teachers lack domain knowledge for a rare drug or a recently approved indication, the resulting pseudo-labels and reference rationales will propagate these blind spots to the student. Developing targeted mechanisms—such as human-in-the-loop calibration for low-confidence cases or retrieval-augmented teacher ensembling to incorporate up-to-date clinical evidence—is a valuable direction for future work.

## Potential Risks

A critical risk in pharmaceutical search relevance is **knowledge staleness**. Drug knowledge graphs and clinical guidelines are continuously updated—new drugs are approved, indications are expanded, and contraindications are revised. If the structured knowledge $\mathcal{K}_p$ or training data $D_{\text{distill}}$ are not regularly synchronized with authoritative sources, the model may rely on outdated information and produce unsafe relevance judgments. This requires operational pipelines for periodic knowledge-base refresh and model re-validation in production environments.

## Ethical Considerations

This work concerns pharmaceutical search relevance, where incorrect recommendations carry health risks. Our system is designed to assist, not replace, qualified medical professionals and human reviewers. Data privacy was handled in accordance with platform policies; no personally identifiable information was used in model training or evaluation. We explicitly note that our relevance labels reflect product-query semantic matching rather than medical advice, and that any health-related decision should be made in consultation with licensed healthcare providers.

## References


Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sunil Arya, David M Mount, Nathan S Netanyahu, Ruth Silverman, and Angela Y Wu. 1998. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM (JACM)*, 45(6):891–923.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 4(5).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Chenhe Dong, Shaowei Yao, Pengkun Jiao, Jianhui Yang, Yiming Jin, Zerui Huang, Xiaojiang Zhou, Dan Ou, Haihong Tang, and Bo Zheng. 2025. Taosr1: The thinking model for e-commerce relevance search. *arXiv preprint arXiv:2508.12365*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2333–2338.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Pengkun Jiao, Yiming Jin, Jianhui Yang, Chenhe Dong, Zerui Huang, Shaowei Yao, Xiaojiang Zhou, Dan Ou, and Haihong Tang. 2025. Taosr-she: Stepwise hybrid examination reinforcement learning framework for e-commerce search relevance. *arXiv preprint arXiv:2510.07972*.

Chenji Lu, Zhuo Chen, Hui Zhao, Zhiyuan Zeng, Gang Zhao, Junjie Ren, Ruicong Xu, Haoran Li, Songyan Liu, Pengjie Wang, and 1 others. 2025. Lore: A large generative model for search relevance. *arXiv preprint arXiv:2512.03025*.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-tuning llama for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2421–2425.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Karsten Schrör. 2007. Aspirin and reye syndrome: a review of the evidence. *Pediatric Drugs*, 9(3):195–204.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. *URL https://arxiv.org/abs/2402.03300*, 2(3):5.

Meituan LongCat Team, Anchun Gui, Bei Li, Bingyang Tao, Bole Zhou, Borun Chen, Chao Zhang, Chengcheng Han, Chenhui Yang, Chi Zhang, and 1 others. 2025. Introducing longcat-flash-thinking: A technical report. *arXiv preprint arXiv:2509.18883*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Jianhui Yang, Yiming Jin, Pengkun Jiao, Chenhe Dong, Zerui Huang, Shaowei Yao, Xiaojiang Zhou, Dan Ou, and Haihong Tang. 2025b. Taosr-agrl: Adaptive guided reinforcement learning framework for e-commerce search relevance. *arXiv preprint arXiv:2510.08048*.

Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: Bert and beyond. In *Proceedings of the 14th ACM International Conference on web search and data mining*, pages 1154–1156.

Dezhi Ye, Jie Liu, Junwei Hu, Jiabin Fan, Bowen Tian, Haijin Liang, and Jin Ma. 2025. Applying large language model for relevance search in tencent. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5171–5181.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, and 1 others. 2025a. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.

Ziyang Zeng, Heming Jing, Jindong Chen, Xiangli Li, Hongyu Liu, Yixuan He, Zhengyu Li, Yige Sun, Zheyong Xie, Yuqing Yang, and 1 others. 2025b. Optimizing generative ranking relevance via reinforcement learning in xiaohongshu search. *arXiv preprint arXiv:2512.00968*.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, and 1 others. 2025. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*.

Lixin Zou, Shengqiang Zhang, Hengyi Cai, Dehong Ma, Suqi Cheng, Shuaiqiang Wang, Daiting Shi, Zhicong Cheng, and Dawei Yin. 2021. Pre-trained language model based ranking in baidu search. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 4014–4022.

# Appendix

The following material is supplementary and does not count toward the 8-page main text limit.

## A Dataset Statistics

Table 5 summarizes the class distribution of the training sets.

Table 5: Statistics of datasets

| Datasets | Highly rel. | Partial rel. | Irrel. |
|---|---|---|---|
| $D_{distill}$ | 63.35k | 31.60k | 54.84k |
| $D_{high_risk}$ | 0 | 13.3k | 6.5k |

## B Implementation Details

During the Supervised Fine-Tuning (SFT) phase, we used a batch size of 1024 and a learning rate of $1 \times 10^{-6}$, managed by a cosine scheduler with a warmup ratio of 0.05. The model was trained

for 10 epochs with a maximum input length of 4096 tokens.

In the GRPO reinforcement learning phase, we set the rollout batch size to 64 and used a learning rate of $1 \times 10^{-6}$. For each input sample, we generated 16 responses via stochastic sampling with a temperature of 0.6 and top-$k$ truncation at $k = 100$. Additionally, we employed a reward-based weighting scheme where:

- $\omega = 0.5$ is the weight for difficult examples; and $\omega_h = 0.6$ is the base weight assigned to high-risk samples in the loss;
- $\alpha = 0.9$ is the cosine similarity threshold used to convert semantic similarity between generated rationales and reference representations into a probability score;
- $\beta = 1$ denotes reasoning process and relevance results are equally important.

**Deployment Notes.** Because pharmaceutical products are managed at the UPC-code level, the total number of items is significantly smaller than in general e-commerce, making offline inference with an 8B-scale model feasible. In actual inference, we only output the relevance results and not the reasoning process to reduce latency without reducing performance.

## C Online Evaluation Metrics

For online evaluation, we conduct live A/B experiments and rely on human assessment to compute the following key metrics:

- **PV (Page View)**: The number of real online impressions served to users. We measure the relative change in PV between the test and baseline groups as an indicator of traffic quality and system efficiency, reflecting how the relevance model affects actual user exposure in production.
- **CTR (Click-Through Rate)**: Computed as $\frac{\text{Clicks}}{\text{Impressions}} \times 100\%$, this metric reflects the relevance and user engagement of an advertisement or search result. A higher CTR typically indicates stronger alignment between user intent and the displayed content.

## D LLM Confidence Scores

Table 6 reports the confidence and corresponding F1 scores of different LLMs with the proposed knowledge-enhanced prompt on the RANDOM evaluation set (here confidence serves as a proxy for F1).

Table 6: Confidence and F1 scores of different LLMs with the proposed knowledge-enhanced prompt.

| Model | Conf./F1 (%) |
|---|---|
| `Qwen3-235B-A22B-Thinking*` | 83.43 |
| `Qwen3-235B-A22B-Instruct` | 82.87 |
| `Qwen3-Max` | 81.25 |
| `LongCat-Large-Thinking` | 82.55 |
| `glm-4.5v` | 81.71 |
| **Confidence-weighted voting** | 84.61 |

## E Prompt Design Ablations

Table 7 compares different prompt strategies for Qwen3-8B.

Table 7: Performance comparison of different prompt strategies for Qwen3-8B.

| Prompt strategy | F1(3-class) | F1(2-class) |
|---|---|---|
| $\Pi(q, p)$ | 56.13% | 85.34% |
| $\Pi(q, p; \mathcal{G})$ | 64.83% | 84.62% |
| $\Pi(q, p, \mathcal{K}_p)$ | 72.93% | 87.65% |
| $\Pi(q, p, \mathcal{K}_p; \mathcal{G})$ | 74.82% | 86.58% |

## F Reward Design Analysis

As discussed in Section 4.3, our sparse binary reward outperforms both dense similarity signals and explicit stepwise matching. We provide additional analysis below.

**Multi-stage training with mixed data prevents catastrophic forgetting.** A third GRPO stage trained exclusively on $D_{\text{high_risk}}$ causes significant performance degradation on general queries (catastrophic forgetting). In contrast, mixing $D_{\text{distill}}$ and $D_{\text{high_risk}}$ during this stage preserves general capabilities while improving performance on critical edge cases. Compared with using mixed data in RL-process rather than two stage RL, it still shows limited performance on high-risk samples, proving that RL based on the first stage can provide a better reasoning capabilities and further improve its performance in another RL stage.

**Impact of Data Distribution.** We observe that simply mixing $D_{\text{high_risk}}$ into the main RL stage (RL-Process) yields a slight degradation in 3-class F1 (86.10% $\rightarrow$ 85.75%). We attribute this to the distribution mismatch: $D_{\text{high_risk}}$ contains no "highly relevant" samples but concentrates on boundary cases (partially relevant and irrelevant). When mixed, these challenging but skewed samples dilute the model's ability to confidently identify clear positive matches, demonstrating that high-risk sam-

ples require isolated staged optimization rather than naive data augmentation.

## G Hyperparameter Sensitivity

We evaluate two values of the cosine-similarity threshold $\alpha$ used in the process reward (Eq. 2). Setting $\alpha = 0.8$ yields 77.06% 3-class F1 and 86.41% 2-class F1, whereas raising $\alpha$ to 0.9 improves 3-class F1 to 86.10% (2-class: 87.20%). The drop at $\alpha = 0.8$ confirms that a permissive threshold admits too many low-quality reasoning traces as positive signals, whereas $\alpha = 0.9$ strikes a better balance between signal sparsity and discriminative power.