LoRE: Enhancing Search Relevance with Progressive Chain-of-Thought and Preference Alignment

Anonymous ACL submission

Abstract

E-commerce search relevance is a critical component of retrieval systems. While Large Language Models (LLMs)-driven Chain-of-Thought (CoT) modeling has become the dominant paradigm and yielded significant gains, a critical gap remains: the absence of a systematic definition for comprehensive relevance reasoning, which leads to significant blind spots in current approaches. In this paper, we deconstruct the task into three core competencies: reasoning & knowledge, multi-modal understanding, and rule awareness. Accordingly, we propose LoRE(**L**arge Generative M**o**del for Search **R**elevanc**e**), a novel two-stage training framework. We first employ an SFT phase to instill these capabilities via a progressive CoT synthesis pipeline, followed by a Reinforcement Learning(RL) phase, which serves as a regularizer, pruning redundant logic to achieve precise and robust adjudication. Extensive experiments validate LoRE, outperforming GPT-5 by 29.1% in Macro-F1 and achieving a relative 27% online gain, offering a vital reference for industrial domain-specific post-training.

1 Introduction

Search relevance plays a pivotal role in E-commerce retrieval systems, which assign scores to query-item pairs and filter out irrelevant items (Liu et al., 2022; Carmel et al., 2020; Yuan et al., 2023). Typically formulated as a classification task, relevance assessment necessitates deep domain expertise and strict rule adherence, thereby imposing rigorous demands on a model's knowledge integration and reasoning capabilities. Although generalist Large Language Models (LLMs) like GPT-5 possess powerful emergent abilities, they suffer from an inherent "Context Gap" regarding specialized vertical knowledge and rigid business criteria. To bridge this divide, recent research (Mehrdad et al., 2024; Fang et al., 2025) has pivoted to domain-specific post-training (Herold et al., 2025; Liu et al., 2025; Zhong et al., 2025), demonstrating significant performance breakthroughs.

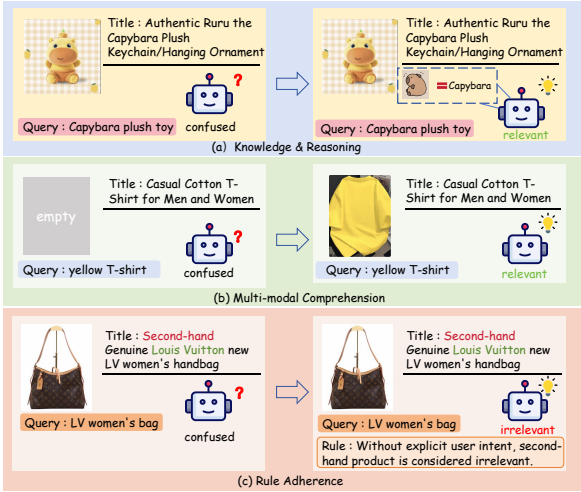


Figure 1: Deconstruction of the relevance task.

Strategies for optimizing LLMs in this specific domain have evolved through two key stages: **(1) Vanilla Supervised Fine-Tuning (SFT) Stage**: Early works like Walmartllm(Mehrdad et al., 2024) and LPFT(Liu et al., 2024) employed direct classification, with the latter enhancing hard negative mining. However, this paradigm struggles with out-of-distribution(OOD) generalization and lacks interpretability. **(2)Chain-of-Thought(CoT) Modeling Stage**: This approach prioritizes explicit CoT reasoning. While ELLM (Zhao et al., 2025a) introduced attribute matching, it lacks fine-grained rule awareness. LREF(Tang et al., 2025) and TaoSR1 (Dong et al., 2025) integrated rule-following but remained constrained to unimodal text, overlooking visual cues. These limitations arise because the literature lacks a clear definition of *what constitutes comprehensive competencies for relevance tasks.*

To address this, we decompose relevance assessment into three core competencies, as summarized in Figure 1: (1) **Knowledge and Reasoning**(Team et al., 2025; Li et al., 2025b; Feng et al., 2024): The model must possess domain knowledge to decipher

specialized terminology and robust reasoning capabilities to infer implicit user intent behind the query, as shown in Figure 1(a). (2) **Multi-modal Understanding**(Gong et al., 2024; Zhang et al., 2024; Li et al., 2024): Since item attributes are dispersed across modalities, the model requires cross-modal alignment to verify query constraints (e.g., color or style) against visual evidence, particularly when textual descriptions are incomplete (Figure 1(b)). (3) **Rule Awareness**: Relevance is often governed by explicit, non-semantic criteria rather than simple similarity. The model must strictly adhere to these predefined business rules (e.g., distinguishing new vs. second-hand items) to ensure accurate adjudication (Figure 1(c)).

To effectively encapsulate these capabilities, we present LoRE (**L**arge Generative M**o**del for Search **R**elevanc**e**), a novel framework that employs a two-stage training regimen. **(1) The SFT stage:** to establish foundational capabilities, serving as a "cold start" for the model. We employ a progressive strategy to synthesize CoT data that encompasses all required competencies by sequentially integrating e-commerce knowledge via RAG, fostering visual reasoning through cross-modal synthesis, and ultimately generating complete reasoning paths grounded in rule guidelines and Teacher LLM inference. Subsequently, we utilize an adaptive data ratio strategy during SFT to effectively inject these reasoning capabilities, thereby elevating the model's performance upper bound. **(2) Reinforcement Learning (RL) stage:** Since SFT tends to mimic superficial patterns rather than mastering intrinsic logic, we employ RL to facilitate exploration. This allows the model to rectify data biases and align with human preferences by suppressing erroneous reasoning. Specifically, we adapt Reinforcement Learning with Verifiable Rewards (RLVR) for relevance assessment, incorporating a curriculum learning(Bengio et al., 2009) strategy based on cold-start sample difficulty. Furthermore, we investigate multiple optimization strategies to mitigate entropy collapse(He et al., 2025a; Yu et al., 2025), which proved effective in stabilizing the training process and boosting performance. Our main contributions can be summarized as follows:

- **Systematic Deconstruction**: We deconstruct the search relevance task into three core competencies: knowledge & reasoning, multi-modal matching, and rule awareness, providing a structured theoretical basis.
- **The LoRE Framework**: We propose LoRE, a two-stage paradigm encompassing the entire pipeline from reasoning data synthesis to model training, which offers a robust reference for domain-specific LLM post-training.
- **Significant Gains**: LoRE demonstrates robust superiority, outperforming GPT-5 by a substantial margin of 29.1% in Macro-F1 on the RAIR benchmark(Lu et al., 2025) while achieving a relative 27% cumulative GoodRate increase in full-scale deployment.

2 Related Work

2.1 E-Commerce Search Relevance

Traditional relevance methods evolved from statistical metrics (e.g., BM25(Robertson and Walker, 1994), TF-IDF(Sparck Jones, 1988)) to deep semantic models, predominantly categorized into representation-based dual-tower architectures(Yang et al., 2025b; Humeau et al., 2019; Huang et al., 2013; Reimers, 2019) and interaction-based single-tower models like BERT(Devlin, 2018; Liu, 2019; Wang et al., 2019). Recently, the field has shifted toward generative reasoning driven by LLMs. Early adaptations(Mehrdad et al., 2024) leveraged LLMs' parametric knowledge to address long-tail queries. Current research focuses on enhancing reasoning and alignment: ELLM(Zhao et al., 2025a) utilizes CoT for explicit attribute extraction, while methods like LREF(Tang et al., 2025) and ProRBP(Chen et al., 2025) synthesize discriminative rules via Direct Preference Optimization (DPO)(Rafailov et al., 2024). Notably, TaoSR1(Dong et al., 2025) further advances this via rule injection and Reinforcement Learning (DeepSeek-AI et al., 2025) but remains text-bound. In contrast, LoRE integrates multi-modal capabilities and employs a progressive pipeline to synthesize a more comprehensive CoT.

2.2 LLM for Classification

E-commerce relevance is formulated as a generative classification task. While RL has revolutionized reasoning domains, its application to classification remains underexplored. Notably, GenCLS++ (He et al., 2025b) observes that, unlike in math tasks, standard CoT often degrades classification performance. Similarly, in multimodal settings, Li et al. (Li et al., 2025a) proposed Adaptive-Thinking to reconcile reasoning with accuracy. Despite these advances, the mechanisms driving RL-induced improvements in generative classification remain poorly understood.

3 Task Definition

3.1 Relevance Task

While the relevance task is fundamentally a binary prediction for a query-item pair (Q, I), we adopt the four-level taxonomy from RAIR (Lu et al., 2025) ($\{L1, L2, L3, L4\}$) to capture granular degrees of satisfaction. In this scheme, the levels are bifurcated such that $L1$–$L2$ represent the *irrelevant* class and $L3$–$L4$ the *relevant* class. Furthermore, we decompose user intent into 18 distinct attribute dimensions (e.g., category, brand) to ensure precise assessment.

3.2 Rule-Aware Relevance Assessment

Practical relevance assessment demands precision that extends beyond general logical reasoning. As is customary in the e-commerce domain, explicit **Rules**(R) are employed not to conflict with general knowledge, but to supplement it by resolving ambiguities. This industry-standard approach transforms subjective interpretations into quantifiable and reproducible metrics. By rigorously defining constraints—such as specifying time windows for "new arrivals"—these rules mitigate human bias and ensure consistent, deterministic adjudication.

Consequently, the relevance assessment task can be formally modeled as predicting a label y conditioned on these applicable rules, rather than relying solely on implicit semantic matching. This formulation is defined as:

$$y = f(Q, I \mid R) \tag{1}$$

where R provides the essential constraints and context to ground the relevance judgment in a consistent, objective, and verifiable manner.

4 Method

4.1 Overview

Section 4.2 details how we progressively synthesize CoT data and inject these reasoning capabilities into the model via SFT with adaptive data ratios. To assess the sufficiency of the cold start in raising the performance ceiling, we employ pass@8. As formulated in Eq. 2, this metric measures the correctness of the binary classification outcome.

$$pass@8 = 1 - \prod_{i=1}^{8} (1 - p_i) \tag{2}$$

Section 4.3 introduces verifiable outcome rewards tailored for relevance tasks, aligning the model with human preferences while suppressing erroneous reasoning. Here, we focus on pass@1, aiming to convert the generative potential of pass@8 into precise, first-attempt accuracy.

4.2 SFT: Comprehensive Reasoning Capability Injection

4.2.1 Multi-dimensional CoT Synthesis

To endow the model with comprehensive and in-depth reasoning capabilities, we propose a progressive CoT synthesis pipeline, as illustrated in Figure 2. This pipeline encompasses three progressive stages: (1) Query Understanding based on Knowledge Injection & Reasoning, (2) Item Understanding based on Multimodal Comprehension, and (3) Relevance Assessment based on Rule Awareness.

Step 1: Query Understanding based on Knowledge Injection and Reasoning While current LLMs possess extensive general world knowledge, they often lack the specific e-commerce expertise required for accurate query understanding. Consequently, they struggle with domain-specific entities where literal interpretation leads to ambiguity—for instance, misclassifying the book title "Incredible Things Also Have Lovers" as "novelty toys". Without such domain priors, precise relevance assessment remains unattainable.

To bridge this knowledge gap, we implement a Retrieval-Augmented Generation (RAG) mechanism anchored by a dynamic e-commerce database. Specifically, we enhance **Query-side** understanding by retrieving titles of high-Click-Through Rate (CTR) items. Recognizing that click-based signals may introduce noise, we employ the Qwen3-235B model to filter these candidates semantically. We retain 5 relevant item titles(excluding the target item itself to prevent leakage), thereby minimizing noise while effectively aggregating the intent behind long-tail queries. Complementarily, we enrich **Item-side** context using merchant-supplied "selling points," providing refined summaries of core attributes and competitive advantages. Let K_Q and K_I denote the retrieved knowledge; the augmented context is formulated as:

$$Context_i = Concat(K_Q, K_I) \tag{3}$$

Building on the retrieved knowledge, the model leverages reasoning to translate query requirements—particularly implicit ones—into concrete attribute constraints. For instance, "summer clothes" implies unstated attributes like "breathable". We prompt the teacher model (Qwen3-235B-

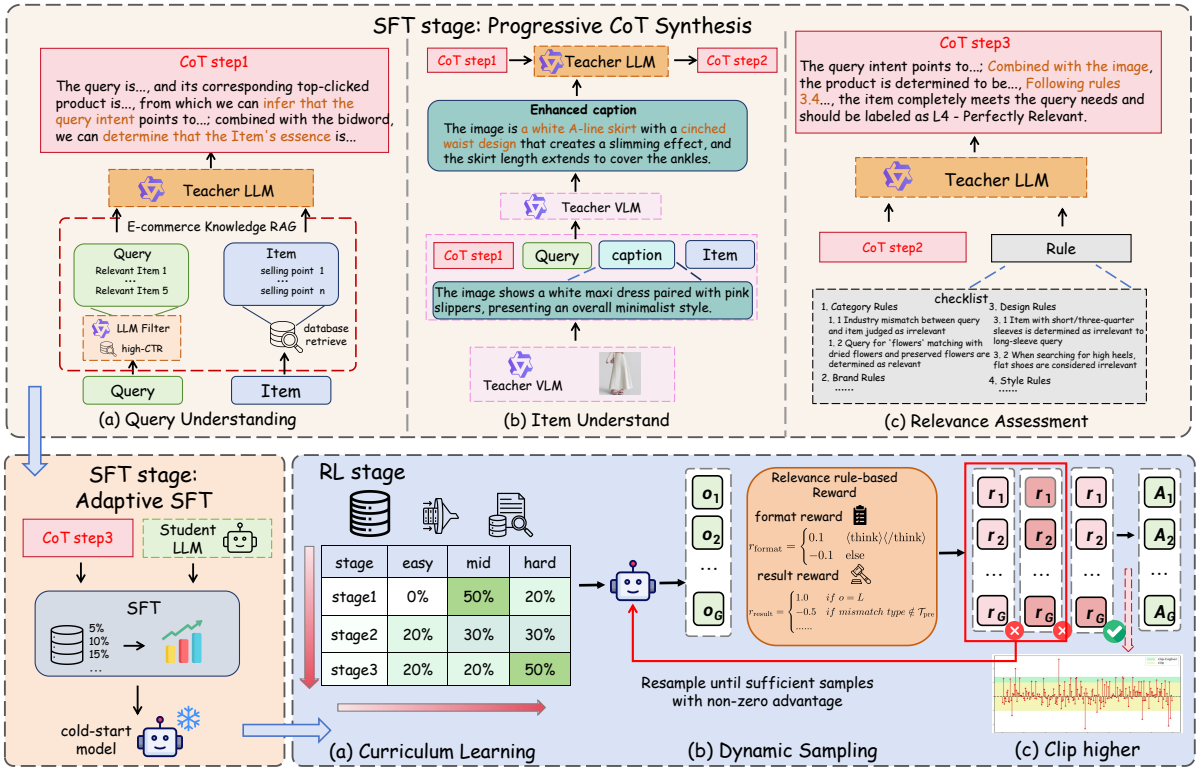


Figure 2: The overview of LoRE Framework.

Instruct) with the augmented context to perform inference, synthesizing internal and domain knowledge to generate the first reasoning step, CoT_{step1}:

$$\text{CoT}_{step1} = \arg\max_y P_\theta(y \mid Concat(Q_i, I_i, context_i)) \tag{4}$$

Step 2: Item Understanding based on Multimodal Comprehension Given that item attributes are presented across both textual and visual modalities, the model must extract and match features from these multi-modal sources after parsing query requirements. We adopt a two-stage caption-mediated framework over direct VLM usage to mitigate reasoning limitations(see Appendix A.1). Specifically, a VLM extracts visual semantics as textual descriptions, which are then integrated into the LLM prompt.

However, generic captioning often omits task-critical visual cues due to the lack of specific guidance. To address this, we propose a relevance-guided multimodal CoT generation strategy: First, we prompt Qwen2.5-VL-72B to generate a basic image caption, denoted as C_{naive}, where T_{VL} represents a carefully crafted prompt template designed to direct the model's attention to salient visual elements in the image.

$$C_{naive} = \arg\max_y P_\theta(y \mid T_{VL}, I_i) \tag{5}$$

Second, to focus on task-relevant details, we integrate the initial caption C_{naive}, the attribute analysis from CoT_{step1} and other context into a structured prompt. This guides Qwen2.5-VL-72B to regenerate a refined caption $C_{enhanced}$, which targets the specific visual attributes required by the query:

$$C_{enhanced} = \arg\max_y P_\theta\big(y \mid \text{Concat}(T_{VL}, Context_i, CoT_{step1}, C_{naive})\big) \tag{6}$$

Finally, by injecting $C_{enhanced}$, CoT_{step1} and contextual information into Qwen3-235B-Instruct, we generate the CoT_{step2}, which effectively encapsulates a holistic understanding of the query intent and item characteristics.

$$\text{CoT}_{step2} = \arg\max_y P_\theta\big(y \mid \text{Concat}(Q_i, I_i, \text{Context}_i, CoT_{step1}, C_{\text{enhanced}})\big) \tag{7}$$

Step 3: Relevance Assessment based on Rule Awareness. While CoT_{step2} bridges the semantic gap, it lacks the rigid verification mandated by Section 3.2. To enforce rigorous attribute matching, we embed rule constraints directly into the reasoning steps, allowing the model to internalize

criteria implicitly via demonstration. Specifically, utilizing industry-specific rule sets R, we construct a prompt containing R, $context_i$, ground truth L_i , and the prior CoT_{step2}. This guides the model to reverse-engineer a rule-compliant reasoning chain CoT_{step3}, which integrates the explicit rule framework with deep semantic comprehension.

$$CoT_{step3} = \arg\max_{y} P_\theta\big(y \mid \text{Concat}(Q_i, I_i, Context_i, CoT_{step2}, L_i, R)\big) \tag{8}$$

4.2.2 Adaptive Distillation by SFT

SFT samples To distill these multi-dimensional reasoning capabilities into the base model, we SFT using the synthesized CoT_{step3} as training targets, as formulated in Eq. 9. To facilitate efficient result extraction and the subsequent RLVR optimization, we structure the response data to strictly separate the reasoning chain from the final conclusion as shown in Eq. 10

$$\mathcal{L}_{\text{SFT}} = -\sum_{i=1}^{N} \log P(y_i|x_i;\theta) \tag{9}$$

$$\text{response} = \texttt{<think></think> <answer></answer>} \tag{10}$$

Prompt Design To optimize inference efficiency, the SFT prompt excludes the exhaustive rule set R used during data synthesis, retaining only essential instructions, which compels the model to internalize the adjudication logic into its parameters rather than relying on explicit context, thereby significantly reducing inference latency. Detailed templates are provided in Appendix D.

Adaptive Ratio A critical challenge is determining the optimal data scale to balance knowledge acquisition against overfitting. To address this, we employ a sensitivity analysis by incrementally increasing the training data in 10% intervals. Throughout this process, we monitor Format Accuracy (reflecting reasoning pattern mastery) and Pass@8 (indicating the performance ceiling). We adopt an adaptive strategy where the data scale is expanded until both metrics reach saturation, ensuring the model maximizes its potential without performance degradation(detailed in Appendix A.2).

4.3 RL: Relevance-Oriented Human Preference Alignment

4.3.1 Basic RLVR Framework

Label-Verified Outcome Reward We introduce RL to guide the model to align with human preferences. We designed a task-specific reward system (Eq. 11, 12, 13) comprising two components: a format reward and an outcome reward. Here, o denotes the model output, L represents the ground truth across levels L1–L4, BinaryClass $B()$ indicates the binary relevance corresponding to these levels, and $\mathcal{T}_{pre}$ refers to the 18 types of attribute deficiency mismatches defined in Section3.1. t_{mis} denotes the mismatch type identified by the model and extracted via regular expressions.

$$r = r_{\text{format}} + r_{\text{result}} \tag{11}$$

$$r_{\text{format}} = \begin{cases} 0.1 & o \in \text{Eq.9} \\ -0.1 & \text{otherwise} \end{cases} \tag{12}$$

$$r_{\text{result}} = \begin{cases} 1.0 & o = L \\ 0.3 & o \neq L \wedge B(o) = B(L) \wedge t_{\text{mis}} \in \mathcal{T}_{\text{pre}} \\ -0.5 & o \neq L \wedge B(o) = B(L) \wedge t_{\text{mis}} \notin \mathcal{T}_{\text{pre}} \\ -1.0 & o \neq L \wedge B(o) \neq B(L) \\ -1.0 & o \text{ is unparsable} \end{cases} \tag{13}$$

The proposed reward mechanism addresses two objectives: (1) enforcing an Output Constraint by strictly limiting predictions to the predefined label set to prevent unconstrained generation; and (2) mitigating reward sparsity via a Hierarchical Shaping Reward. By assigning partial credit (e.g., 0.3) to predictions that capture binary relevance despite fine-grained mismatches, it encourages the model to prioritize the fundamental decision boundary, thereby accelerating early-stage convergence.

GRPO Algorithm Figure 2 illustrates our RL framework utilizing the GRPO algorithm. We optimize the objective function (Eq. 14) to maximize the advantage-weighted log probability expectation, with the importance sampling ratio and advantage function defined in Eq. 15 and Eq. 16. We discard KL constraints to facilitate unencumbered exploration. Crucially, to prevent reward hacking, the format reward (Eq. 12) acts as a structural anchor, which effectively functions as a proxy for KL regularization, guaranteeing valid output while allowing model to freely discover optimal paths.

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x,y}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} \min\Big(w_t\widehat{A}_t, \text{clip}(w_t, 1-\varepsilon, 1+\varepsilon)\widehat{A}_t\Big)\Bigg] \tag{14}$$

$$w_{i,t} = \frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})} \tag{15}$$

$$\widehat{A}_{i,t} = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^G)}{\text{std}(\{r_j\}_{j=1}^G)} \tag{16}$$

4.3.2 Training Strategy Optimization

Curriculum Learning We introduce a curriculum learning strategy in RL training. To assess sample difficulty, we performed 8 inference runs per sample using the cold-start model at a set temperature. After excluding trivial cases (where all predictions were correct), we classified the remaining samples as follows:

$$\text{Difficulty}(x) = \begin{cases} \text{Easy} & \text{if } k(x) \in \{6, 7\} \\ \text{Medium} & \text{if } k(x) \in \{3, 4, 5\} \\ \text{Hard} & \text{if } k(x) \in \{0, 1, 2\} \end{cases} \tag{17}$$

Subsequently, we implement a three-stage curriculum learning scheme with progressive difficulty, as detailed in Table 1.

Table 1: Curriculum learning sample allocation. Percentages denote the proportion of each difficulty pool used; samples are non-overlapping across stages.

stage	easy	mid	hard
stage 1	0%	50%	20%
stage 2	20%	30%	30%
stage 3	20%	20%	50%

The curriculum design follows three principles: (1) Rapid Bootstrap, where Stage 1 targets medium-difficulty instances to balance feedback and challenge; (2) Progressive Challenge, increasing difficulty in Stages 2–3 to extend capabilities to edge cases; and (3) Knowledge Retention, which incorporates easy samples to prevent forgetting and stabilize training. We empirically determined the optimal data proportions (Table 1) through grid search.

Dynamic Sampling To mitigate efficiency decay caused by saturated samples (where $A_i = 0$ due to consistent outcomes), we introduce a dynamic sampling strategy(Yu et al., 2025). This method discards zero-advantage samples during rollout and triggers policy updates only when the accumulated number of effective samples meets the batch size, thereby optimizing computational resources.

Clip-higher The rapid decrease in entropy during training restricts the discovery of superior reasoning paths. We explored several methods(He et al., 2025a)(Section 5.5 to slow this decline and preserve exploration space. Among these, Clip-higher(Yu et al., 2025) proved to be the optimal strategy, effectively countering premature stabilization and yielding the best performance gains.

5 Experiment

5.1 Setting

Evaluation Datasets and Metrics We evaluate our model on RAIR(Lu et al., 2025), a challenging e-commerce dataset derived from Taobao, which is stratified into General, Hard, and Visual Salient subsets, and its explicit taxonomy serves as the basis for our model design. We employ three metrics for comprehensive evaluation:

Binary Accuracy (Acc@2): Groups L1/L2 as irrelevant and L3/L4 as relevant to evaluate binary relevance detection capabilities. Let $B(y)$ map $\{L1, L2\} \to 0$ and $\{L3, L4\} \to 1$:

$$\text{Acc@2} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(B(\hat{y}_i) = B(y_i)) \tag{18}$$

Fine-grained Accuracy (Acc@4): Measures the standard classification accuracy across the four specific relevance levels (L1–L4):

$$\text{Acc@4} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = y_i) \tag{19}$$

Macro-F1: Calculated as the arithmetic mean of per-class F1 scores to address label imbalance, where K is the number of classes:

$$\text{Macro-F1} = \frac{1}{K}\sum_{k=1}^{K} \text{F1}_k \tag{20}$$

Baselines We establish two baselines: (1) Top-tier LLMs, evaluated in a zero-shot setting using official parameters and prompts incorporating the rules R; and (2) Vanilla SFT, trained without CoT or RL to isolate our framework's effectiveness.

5.2 Offline Evaluation

Both Vanilla SFT and LoRE utilize the Qwen2.5-7B backbone (Appendix C.1) and are evaluated using greedy decoding. We exclude target and its textually identical items from the RAG pool during inference to prevent leakage, ensuring reliance on reasoning rather than memorization.

Tables 2 and 3 present our main experimental findings, summarized as follows:

1. Domain-specific SFT yields substantial gains. SFT models significantly outperform baselines, including closed-source SOTA models like

Table 2: Main results on the General subset and Long-Tail hard subset. Bold indicates the best performance. LoRE achieves significant gains, particularly on the Hard subset.

Model	General subset			Long-Tail Hard subset		
	Acc@2	Acc@4	Macro-F1	Acc@2	Acc@4	Macro-F1
prompting-based models						
Qwen2.5-7B(base)(Qwen et al., 2025)	0.775	0.689	0.395	0.531	0.382	0.312
Qwen3-30B-Instruct(Yang et al., 2025a)	0.763	0.683	0.391	0.587	0.431	0.333
Qwen3-235B-Instruct(Yang et al., 2025a)	0.830	0.676	0.417	0.609	0.381	0.359
Llama3.1-8B-Instruct(Grattafiori et al., 2024)	0.788	0.416	0.224	0.525	0.248	0.200
Llama3.1-70B-Instruct(Grattafiori et al., 2024)	0.774	0.668	0.369	0.547	0.378	0.295
Qwen3-4B-Thinking(Yang et al., 2025a)	0.805	0.720	0.453	0.584	0.419	0.357
Qwen3-30B-Thinking(Yang et al., 2025a)	0.784	0.718	0.457	0.582	0.448	0.362
Qwen3-235B-Thinking(Yang et al., 2025a)	0.779	0.702	0.470	0.585	0.451	0.367
Gemini 2.5 Pro(Comanici et al., 2025)	0.795	0.701	0.483	0.627	0.481	0.392
GPT-5	0.845	0.714	0.433	0.681	0.435	0.407
vanilla SFT	0.929	0.891	0.722	0.671	0.542	0.413
LoRE	**0.933**	**0.897**	**0.724**	**0.715**	**0.582**	**0.460**

Table 3: Results on the Visual Salient subset.

Model	Visual salience subset		
	Acc@2	Acc@4	Macro-F1
prompting-based models			
Qwen2.5-VL-7B-Instruct(Bai et al., 2025)	0.535	0.285	0.230
Qwen2.5-VL-32B-Instruct(Bai et al., 2025)	0.647	0.467	0.339
Qwen2.5-VL-72B-Instruct(Bai et al., 2025)	0.608	0.420	0.267
Gemini 2.5 Pro(Comanici et al., 2025)	0.670	0.561	0.377
GPT-5	0.682	0.508	0.369
Vanilla SFT	0.638	0.574	0.378
LoRE	**0.698**	**0.627**	**0.426**

GPT-5, across both General and Hard subsets. This is most pronounced on the General subset, where LoRE surpasses GPT-5 by 8.8% in Acc@2 and 29.1% in Macro-F1, proving that CoT rule injection outperforms explicit prompting, while also confirming effective domain knowledge acquisition.

2. CoT modeling refines high-performance baselines. On the General subset, the Vanilla-SFT baseline already achieves saturation (0.929 Acc@2), significantly outperforming GPT-5. LoRE yields a further marginal gain of 0.4%, indicating that while SFT suffices for routine in-distribution data, LoRE provides incremental improvements on top of this strong baseline.

3. Reasoning injection boosts performance on Hard samples. Unlike on the General subset, Vanilla-SFT struggles with the Long-Tail Hard subset, showing negligible advantages over GPT-5 (0.671 vs. 0.681 Acc@2) due to limited reasoning capabilities. In contrast, LoRE achieves a breakthrough with an Acc@2 of 0.715 (+4.4%) and Macro-F1 of 0.460 (+5.3%), surpassing strong models like GPT-5. This confirms that the synergy between synthetic CoT and RL amplifies discriminative power on long-tail, challenging scenarios.

4. Relevance-guided captions enhance multimodal reasoning. As shown in Table 3, Vanilla-SFT lacks advantages over advanced VLMs on visual samples. However, LoRE's visual relevance enhancement strategy achieves a 4.8% relative improvement in Macro-F1 over GPT-5, proving that integrating multimodal information into CoT significantly boosts cross-modal understanding.

5.3 Online A/B Testing

We conducted a full-scale deployment on the Taobao platform to validate based on LoRE via three strategies: (1) Cache Deployment: Pre-computes scores for high-frequency query-item pairs using LoRE. (2) Knowledge Distillation(Appendix C.2): Distills LoRE into the online lightweight model using 60M generated samples. (3) System Strategy Update: leverage LoRE scores in downstream ranking, balancing objectives through Pareto optimization. We report GoodRate (the proportion of relevant items displayed). As shown in Table 4, these approaches achieved a cumulative relative +27.0% improvement.

Table 4: Online A/B testing results on Taobao platform.

Strategy	GoodRate
Cache Deployment	+4.8%
Knowledge Distillation (60M)	+9.5%
System & Strategy Update	+12.7%
Overall	**+27.0%**

5.4 Ablation Study

As shown in Table 5, our stepwise ablation validates each component. The Base model employs

vanilla SFT on reasoning paths derived from the query, item, and labels by a teacher LLM.

SFT Insights: The effectiveness of our cross-modal CoT synthesis is evidenced by the gain from *Enhanced Caption*. Crucially, *Rules Integration* provides the most significant boost, especially in Macro-F1 (+9.0%). This indicates that without explicit rules, the model struggles with fine-grained boundaries, whereas rule injection effectively sharpens these distinctions. **RL Insights:** Naive GRPO proves insufficient for substantial gains over the strong cold-start model. It is the introduction of advanced strategies—specifically Curriculum Learning and clip-higher—that unlocks the full potential of RL, resulting in the final robust performance.

Table 5: Ablation study on the General subset.

Method	General Subset		
	Acc@2	Acc@4	Macro-F1
Phase 1: SFT			
Base (Vanilla CoT)	0.815	0.738	0.501
+ RAG	0.823	0.741	0.507
+ Naive Caption	0.828	0.752	0.513
+ Enhanced Caption	0.834	0.759	0.521
+ Rules Integration (**cold-start**)	**0.887**	**0.810**	**0.611**
Phase 2: RL (from cold-start)			
+ GRPO	0.904	0.852	0.657
+ Curriculum Learning	0.921	0.871	0.674
+ Dynamic Sampling	0.926	0.885	0.693
+ clip-higher (**LoRE**)	**0.933**	**0.897**	**0.724**

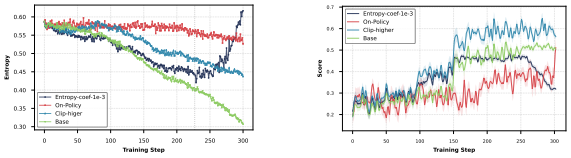
5.5 Analysis and Discussion

Impact of CoT Distillation Paradoxically, while CoT distillation expands the capability upper bound (Pass@8), it degrades single-pass reliability (Pass@1) (Table 6). We attribute this instability to: (1) exposure bias from training-inference discrepancy; and (2) a tendency to "over-reason" on simple queries (Appendix A.3). RL proves indispensable here as a logic pruning mechanism: by penalizing incorrect outcomes, it implicitly suppresses the error-prone redundancy that leads to them, effectively collapsing reasoning chains to their minimal robust paths.

Table 6: Performance comparison between vanilla SFT and cold-start model with CoT distillation on the General subset.

Models	pass@1	pass@8
vanilla SFT	0.929	0.937
cold-start	0.887(-4.2%)	0.964(+2.7%)

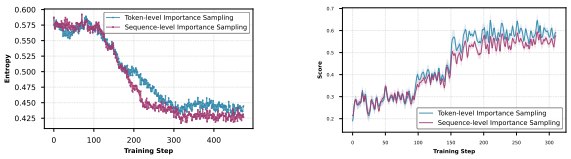
Entropy Collapse Optimization: A Comparison To mitigate entropy collapse, we compared clip-higher against two alternatives: (1) On-policy, which performs a single gradient update per sampled batch to enforce strictly fresh exploration; and (2) Explicit Entropy Regularization, which adds a penalty term to the loss to directly counteract entropy decay. As shown in Figure 3, clip-higher achieves the optimal trade-off. It moderates the rate of entropy decline rather than halting it, ensuring stable convergence alongside sufficient exploration. In contrast, the On-policy approach maintains excessive entropy, hindering convergence, while Entropy Regularization suffers from instability, as the penalty term tends to dominate the loss in later stages, leading to training collapse.



(a) Entropy comparison (b) Reward comparison

Figure 3: Experimental comparison of three entropy collapse mitigation strategies

Importance Sampling Granularity Exploration We compared token-level importance sampling against the sentence-level approach (Zheng et al., 2025; Zhao et al., 2025b) that applies unified sequence weights. As shown in Figure 4, although sentence-level sampling stabilizes rewards, it accelerates entropy collapse. For structured relevance tasks, this leads to premature convergence and restricted exploration. Consequently, the finer-grained token-level sampling yields superior performance.



(a) Entropy comparison (b) Reward comparison

Figure 4: Experimental comparison between token-level and sequence-level importance sampling.

6 Conclusion

In this work, we deconstruct relevance modeling into LoRE, a framework integrating reasoning injection (SFT) and preference alignment (RL). Our findings suggest that for vertical domains, internalizing expert rules via progressive CoT and verifiable RL constitutes a superior paradigm compared to generic large-scale prompting. Extensive experiments validate this approach, demonstrating LoRE's robust superiority over existing baselines.

Limitations

Despite the promising results, this study relies primarily on outcome-based rewards derived solely from the final relevance labels. This sparse reward signal presents a limitation in granularity: it treats the entire reasoning chain as a black box, potentially reinforcing erroneous reasoning paths when they coincidentally lead to the correct prediction (i.e., the "right answer for the wrong reason" phenomenon). In reality, the relevance assessment process can be explicitly decomposed into three distinct and verifiable stages: Query Understanding, Item Understanding, and Relevance Judgment. Since each of these sub-steps possesses deterministic criteria for evaluation, there is an opportunity for finer-grained supervision. In future work, we intend to explore Process Reward Models (PRMs) to provide step-by-step feedback, thereby mitigating the accumulation of reasoning errors and further enhancing the model's interpretability and robustness.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *ICML*.

David Carmel, Elad Haramaty, Arnon Lazerson, Liane Lewin-Eytan, and Yoelle Maarek. 2020. Why do people buy seemingly irrelevant items in voice product search? on the relation between product relevance and customer satisfaction in ecommerce. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, page 79–87, New York, NY, USA. Association for Computing Machinery.

Zeyuan Chen, Haiyan Wu, Kaixin Wu, Wei Chen, Mingjie Zhong, Jia Xu, Zhongyi Liu, and Wei Zhang. 2025. Towards boosting LLMs-driven relevance modeling with progressive retrieved behavior-augmented prompting. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 784–793, Abu Dhabi, UAE. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, and Noveen Sachdeva. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, and Ruoyu Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Zheng Fang, Donghao Xie, Ming Pang, Chunyuan Yuan, Xue Jiang, Changping Peng, Zhangang Lin, and Zheng Luo. 2025. Adore: Autonomous domain-oriented relevance engine for e-commerce. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 4259–4263, New York, NY, USA. Association for Computing Machinery.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690, Bangkok, Thailand. Association for Computational Linguistics.

Yunye Gong, Robik Shrestha, Jared Claypoole, Michael Cogswell, Arijit Ray, Christopher Kanan, and Ajay Divakaran. 2024. BloomVQA: Assessing hierarchical multi-modal comprehension. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14905–14918, Bangkok, Thailand. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and Ahmad Al-Dahle. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, Siyuan Li, Liang Zeng, Tianwen Wei, Cheng Cheng, Bo An, Yang Liu, and Yahui Zhou. 2025a. Skywork open reasoner 1 technical report. *Preprint*, arXiv:2505.22312.

Mingqian He, Fei Zhao, Chonggang Lu, Ziyan Liu, Yue Wang, and Haofu Qian. 2025b. Gencls++: Pushing the boundaries of generative classification in llms through comprehensive sft and rl studies across diverse datasets. *Preprint*, arXiv:2504.19898.

Christian Herold, Michael Kozielski, Tala Bazazo, Pavel Petrushkov, Yannick Versley, Seyyed Hadi Hashemi, Patrycja Cieplicka, Dominika Basaj, and Shahram Khadivi. 2025. Domain adaptation of foundation

LLMs for e-commerce. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1039–1049, Vienna, Austria. Association for Computational Linguistics.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, and 1 others. 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.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387, Bangkok, Thailand. Association for Computational Linguistics.

Ming Li, Jike Zhong, Shitian Zhao, Yuxiang Lai, Haoquan Zhang, Wang Bill Zhu, and Kaipeng Zhang. 2025a. Think or not think: A study of explicit thinking in rule-based visual reinforcement fine-tuning. *Preprint*, arXiv:2503.16188.

Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, Tat-Seng Chua, and Yang Deng. 2025b. Knowledge boundary of large language models: A survey. *Preprint*, arXiv:2412.12472.

Hong Liu, Saisai Gong, Yixin Ji, Kaixin Wu, Jia Xu, and Jinjie Gu. 2024. Boosting llm-based relevance modeling with distribution-aware robust learning. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 4718–4725, New York, NY, USA. Association for Computing Machinery.

Huanghai Liu, Quzhe Huang, Qingjing Chen, Yiran Hu, Jiayu Ma, Yun Liu, Weixing Shen, and Yansong Feng. 2025. JUREX-4E: Juridical expert-annotated four-element knowledge base for legal reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3794–3814, Suzhou, China. Association for Computational Linguistics.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ziyang Liu, Chaokun Wang, Hao Feng, Lingfei Wu, and Liqun Yang. 2022. Knowledge distillation based contextual relevance matching for e-commerce product search. *Preprint*, arXiv:2210.01701.

Chenji Lu, Zhuo Chen, Hui Zhao, Zhenyi Wang, Pengjie Wang, Jian Xu, and Bo Zheng. 2025. Rair: A rule-aware benchmark uniting challenging long-tail and visual salience subset for e-commerce relevance assessment. *Preprint*, arXiv:2512.24943.

Navid Mehrdad, Hrushikesh Mohapatra, Mossaab Bagdouri, Prijith Chandran, Alessandro Magnani, Xunfan Cai, Ajit Puthenputhussery, Sachin Yadav, Tony Lee, ChengXiang Zhai, and Ciya Liao. 2024. Large language models for relevance judgment in product search. *Preprint*, arXiv:2406.00247.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Stephen E. Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Karen Sparck Jones. 1988. *A statistical interpretation of term specificity and its application in retrieval*, page 132–142. Taylor Graham Publishing, GBR.

Tian Tang, Zhixing Tian, Zhenyu Zhu, Chenyang Wang, Haiqing Hu, Guoyu Tang, Lin Liu, and Sulong Xu. 2025. Lref: A novel llm-based relevance framework for e-commerce search. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, page 468–475, New York, NY, USA. Association for Computing Machinery.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, and Chenjun Xiao. 2025. Kimi k1.5: Scaling reinforcement learning with llms. *Preprint*, arXiv:2501.12599.

Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, and Luo Si. 2019. Structbert: Incorporating language structures into pre-training for deep language understanding. *arXiv preprint arXiv:1908.04577*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, and Bo Zheng. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Haoqiang Yang, Congde Yuan, Kun Bai, Mengzhuo Guo, Wei Yang, and Chao Zhou. 2025b. Hit model: A hierarchical interaction-enhanced two-tower model for pre-ranking systems. *Preprint*, arXiv:2505.19849.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.

Chunyuan Yuan, Yiming Qiu, Mingming Li, Haiqing Hu, Songlin Wang, and Sulong Xu. 2023. A multi-granularity matching attention network for query intent classification in e-commerce retrieval. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23, page 416–420. ACM.

Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. 2024. Earthgpt: A universal multimodal large language model for multisensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20.

Gang Zhao, Ximing Zhang, Chenji Lu, Hui Zhao, Tianshu Wu, Pengjie Wang, Jian Xu, and Bo Zheng. 2025a. Explainable llm-driven multi-dimensional distillation for e-commerce relevance learning. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, page 631–640, New York, NY, USA. Association for Computing Machinery.

Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Xun Wu, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, Fang Wan, and Furu Wei. 2025b. Geometric-mean policy optimization. *Preprint*, arXiv:2507.20673.

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

Qihuang Zhong, Liang Ding, Xiantao Cai, Juhua Liu, Bo Du, and Dacheng Tao. 2025. KaFT: Knowledge-aware fine-tuning for boosting LLMs' domain-specific question-answering performance. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24085–24100, Vienna, Austria. Association for Computational Linguistics.

A Extending Quantitative Analysis

A.1 Multimodal Modeling: VLM or Two-Stage LLM?

A more intuitive approach to modeling multi-modal capabilities for relevance tasks is to directly leverage VLMs, as they inherently possess image understanding capabilities without requiring additional caption synthesis. We conducted exploratory experiments on this approach: we employ the teacher model Qwen2.5-VL-72B to sequentially complete two steps—knowledge and reasoning CoT generation, followed by rule-aware CoT generation—thereby synthesizing multi-dimensional CoT data. Subsequently, using the same two-stage training configuration (SFT and RL), we trained the Qwen2.5-VL-7B model, which has a comparable parameter to the LLM base model employed in this study.

Table 7: Performance comparison of VLM-based and LLM-based methods on the general subset and Long-Tail Hard subset.

Model	General subset			Long-Tail Hard subset		
	Acc@2	Acc@4	F1	Acc@2	Acc@4	F1
VLM-base	0.912	0.878	0.670	0.692	0.577	0.413
LLM-base	**0.933**	**0.897**	**0.724**	**0.715**	**0.582**	**0.460**

Table 8: Performance comparison of VLM-based and LLM-based methods on the Visual salience subset.

Model	Visual Salient subset		
	Acc@2	Acc@4	Macro-F1
VLM-base	**0.703**	**0.645**	**0.526**
LLM-base	0.698	0.627	0.426

Tables 7 and 8 present a performance comparison between VLM-based and LLM-based models. On both General and Hard datasets, the LLM-based method significantly outperforms its VLM counterpart, particularly in Macro-F1 metrics. We attribute this advantage to two factors: (1) LLMs inherently possess stronger reasoning and knowledge integration capabilities; and (2) the relevance-guided captioning acts as an effective visual attention filter, retaining only query-relevant elements while discarding visual noise. This abstraction proves sufficient for most cases where relevance hinges on key visual objects. Conversely, VLMs excel on the Visual Salient subset, aligning with expectations. These cases demand fine-grained, pixel-level visual attention that captions may overlook, whereas VLMs process raw images directly, preserving intricate visual details essential for subtle judgments.

A.2 Scaling law effects in SFT

In Section 4.2.2, we propose a data efficiency analysis by incrementally scaling the training set from 10% upwards. We focus on Format Accuracy and

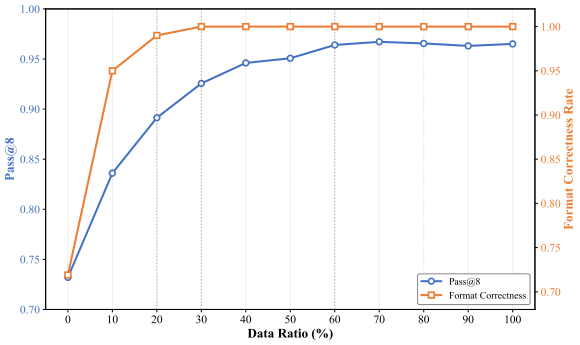


Figure 5: Evolution of pass@8 and format correctness rate with varying data proportions in SFT process.

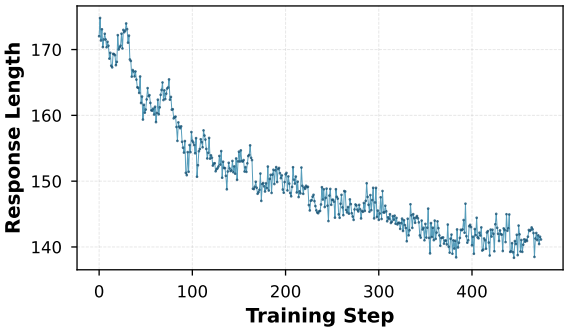


Figure 6: Response length variation during RL training.

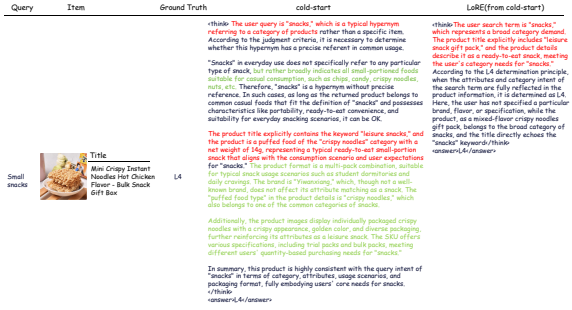


Figure 7: Example analysis of capabilities required for relevance evaluation.

Pass@8 to gauge instruction following and reasoning potential, respectively.

Figure 5 reveals distinct convergence behaviors: format alignment is achieved early (98% accuracy at 20% data), whereas reasoning capability (Pass@8) grows steadily before plateauing. We observe that performance gains become negligible beyond 60% data usage. To maximize training efficiency without compromising capability, we adopt this 60% threshold as the optimal configuration for our cold-start model.

A.3 Long CoT is not necessary for better performance.

Studies such as DeepSeek-R1 have observed the emergent phenomenon of long CoT and the model's "Aha Moment." We similarly conducted a tracking analysis of the dynamic changes in model output length during the reinforcement learning process, with results presented in Figure 6. The experimental findings reveal that as model capability continues to improve, the output length exhibits an overall decreasing trend and eventually stabilizes. This phenomenon suggests that long CoT are not a necessary condition for model capability enhancement, but rather an accompanying phenomenon during the training process. Further analysis indicates that different task types exhibit significant variations in their requirements for reasoning chain length: tasks such as mathematical computation and code generation typically necessitate longer step-by-step reasoning processes. For the relevance task, accurately identifying query intent and item attributes, and reasoning based on discrimination rules, constitutes an intuitive and clear chain of thought. Redundant reasoning processes beyond this do not yield performance improvements.

To further substantiate this, we present a qualitative comparison in Figure 7. The query "Small snacks" represents a straightforward category-based intent. Theoretically, a correct judgment relies solely on identifying the category match (highlighted in red), without requiring complex attribute deduction.

The **cold-start model**, characteristic of early-stage exploration, falls into the trap of "over-reasoning." It expends excessive tokens performing linguistic definitions (e.g., explaining "hypernyms"), analyzing peripheral visual details (e.g., "golden color"), and validating irrelevant attributes like packaging specifications. While the final answer is correct, the reasoning process is filled with noise and redundant verification.

In stark contrast, **LoRE** demonstrates a crystallized reasoning path. It bypasses semantic definitions, directly locks onto the core category alignment (Query: "snacks" → Title: "leisure snack gift pack"), and immediately applies the L4 judgment rule.

This contrast confirms that unlike mathematical computation which necessitates step-by-step expansion, relevance estimation favors an "Occam's Razor" approach: the RL process effectively optimizes the policy by pruning redundant logic. Cru-

cially, this analysis offers a tangible explanation for the Pass@1 degradation observed in naive CoT distillation. The "Cold-start" model illustrates how overly long, redundant reasoning chains introduce noise, thereby increasing the probability of error accumulation. This further validates the necessity of RL: it functions as a critical mechanism to prune superfluous cognitive steps, robustifying the policy against the fragility inherent in verbose reasoning.

B Case Study

From a qualitative perspective, we demonstrate the model's improvements across multiple capability dimensions through representative cases, as shown in Figure 8:

Knowledge. In Case 1, the query term is "Capybara" and the item title is "water pig plush toy." After distillation from the teacher model, the model has internalized sufficient world knowledge to recognize that "Capybara" is an alternative name for "water pig," thereby correctly determining their relevance.

Reasoning. In Case 2, the query is "8 month maternity clothes", while the item title contains no explicit exclusion keywords regarding pregnancy. Through reasoning capability, the model analyzes that: the user's intent for "maternity wear" inherently implies a functional demand for loose, strictly comfortable cuts to accommodate physical changes; conversely, the item emphasizes a "Slim Fit" design, which implies a silhouette structured to hug the body closely. The model deduces that this specific design attribute is functionally incompatible with the spatial and comfort requirements of the target demographic, identifying a hidden mismatch in applicability, leading to a correct judgment.

Multimodal Understanding. Case 3 represents a typical scenario of joint text-image discrimination. For the query "1/4 Black Soft," the model successfully correlates the title attribute "Black Soft" with visual evidence. Crucially, this capability is explicitly attributed to the Relevance-guided Captioning module (Teacher VLM), which accurately extracted the key text "Kase MAGNETIC BLACK MIST 1/4" engraved on the item from the image. By converting this pixel-level signal into explicit text, the pipeline enables LoRE to synthesize visual and textual cues for coordinated reasoning, demonstrating the synergy between the VLM captioner and the reasoning model.

Rule Compliance. In Case 4, the attribute

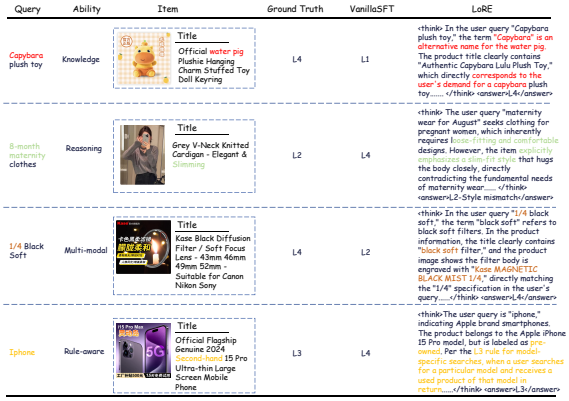


Figure 8: Example analysis of capabilities required for relevance evaluation.

matching between query and item is not inherently complex, but the sample belongs to a special case stipulated by rules: "searching for normal items but returning second-hand items should be judged as L3 (weakly relevant)." After rule-aware CoT modeling, the model has acquired sufficient rule compliance capability to accurately identify and apply such business rules.

C Implementation Details

C.1 Offline Training

We utilize Qwen2.5-7B(Qwen et al., 2025) as the backbone model. For **SFT**, we employ the AdamW optimizer with a learning rate of $1e-5$ and a global batch size of 256. The model is trained on 32 NVIDIA H20 GPUs for approximately 17 hours. For **RL**, we configure the rollout batch size to 256, group size to 8, and sampling temperature to 1.0, upper bound of clip-higher to 0.28. The training batch size is set to 256. This phase requires 39 hours of training on 128 NVIDIA H20 GPUs.

Our training data were collected from Taobao search logs and labeled by professional annotators strictly adhering to the guidelines established in RAIR (Lu et al., 2025). We partitioned the dataset into training and validation sets at a 9:1 ratio. The validation set was specifically utilized to determine the optimal data ratio for the Supervised Fine-Tuning (SFT) stage.

C.2 Online Distillation

To address the architectural distinction between the offline data generation pipeline and the online inference system, we clarify our **asymmetric distillation strategy**. This design balances the trade-off

Prompt: E-commerce Relevance Assessment with CoT

You are an expert in e-commerce relevance analysis. Your task is to assess the relevance between a user **Query** and a **item** based on provided information.

Key Attributes: Pay attention to Category, Brand, Style Details, Material, Target Audience, Season, Model ID, Specifications, Shop, Color, Function, Accessories, Set/Single, Price, Year, Special Attributes, and IP image.
Instructions:

- **Synthesize Information:** Combine insights from the item Title, Shop, Details, Selling Points, and Image Caption.
- **Reference User Intent:** Analyze the **Top Clicked Items** to infer the user's implicit intent (e.g., preferred style or category). **Note: High CTR may stem from popularity rather than relevance; treat these items strictly as auxiliary context to avoid bias.**
- **Leverage Selling Points:** Use **Item Selling Points** to identify specific item highlights or hidden attributes that satisfy the query.
- **Chain of Thought (CoT):** You must enclose your reasoning process within `<think>` and

  </think>

  `</think>`

  `tags before outputting the final result.`
- **`Output Format:`** `For L2 cases, specify the mismatch type in brackets, e.g., [L2-Style Mismatch].`

Relevance Levels:

- **L1 (Irrelevant):** Complete category mismatch with no association.
- **L2 (Partially Irrelevant):** Category mismatch but related; or Category matches but key attributes (e.g., Brand, Spec, Gender) fail.
- **L3 (Closely Relevant):** Proximate category/attributes but lacks full intent alignment; or contains minor attribute conflicts.
- **L4 (Perfectly Relevant):** Completely satisfies the query intent.

Input:
User Query: {query}
Top Clicked Items: {top_click_titles}
Item Title: {title}
Item Shop: {shop_name}
Item Selling Points: {selling_points}
Item Details: {detail_info}
Item SKU: {sku_info}
Image Caption: {image_caption}

Figure 9: The prompt template used for relevance assessment. We explicitly instruct the model to treat high-CTR items as reference-only to mitigate popularity bias.

between maximizing reasoning accuracy and satisfying strict industrial latency constraints.

Offline Teacher (Accuracy-Oriented): During the construction of synthetic training data, our primary goal is to reach the highest possible accuracy upper bound. Therefore, the Teacher model (LoRE) utilizes a resource-intensive configuration: it integrates Relevance-guided VLM Captioning to extract visual details, employs RAG to incorporate top-5 historical high-CTR items for user intent alignment, and processes comprehensive textual metadata (full selling points and detailed descriptions).

Online Student(Latency-Oriented): Conversely, for the online serving stage, real-time access to VLM modules or retrieval engines is computationally prohibitive. Consequently, the deployed student model is a lightweight BERT-like model. It operates on a streamlined input schema consisting solely of the `User Query`, `item Title`, and `Truncated Key Attributes`. Despite lacking direct access to visual and historical context during inference, the student model effectively approximates the teacher's decision boundaries via knowledge distillation.

D Prompt Details

To ensure reproducibility, we provide the prompt used for SFT as shown in Figure 9: