

LREF: A Novel LLM-based Relevance Framework for E-commerce Search

Tian Tang*
tangtian0456@gmail.com
JD.COM
Beijing, China

Zhixing Tian*
tianzhixing2017@gmail.com
JD.COM
Beijing, China

Zhenyu Zhu
zhenzhuyu240@gmail.com
JD.COM
Beijing, China

Chenyang Wang
wangchenyang3@jd.com
JD.COM
Beijing, China

Haiqing Hu
huhaiqing1@jd.com
JD.COM
Beijing, China

Guoyu Tang
tangguoyu@jd.com
JD.COM
Beijing, China

Lin Liu
liulin1@jd.com
JD.COM
Beijing, China

Sulong Xu
xusulong@jd.com
JD.COM
Beijing, China

Abstract

Query and product relevance prediction is a critical component for ensuring a smooth user experience in e-commerce search. Traditional studies mainly focus on BERT-based models to assess the semantic relevance between queries and products. However, the discriminative paradigm and limited knowledge capacity of these approaches restrict their ability to comprehend the relevance between queries and products fully. With the rapid advancement of Large Language Models (LLMs), recent research has begun to explore their application to industrial search systems, as LLMs provide extensive world knowledge and flexible optimization for reasoning processes. Nonetheless, directly leveraging LLMs for relevance prediction tasks introduces new challenges, including a high demand for data quality, the necessity for meticulous optimization of reasoning processes, and an optimistic bias that can result in over-recall.

To overcome the above problems, this paper proposes a novel framework called the LLM-based RElevance Framework (LREF) aimed at enhancing e-commerce search relevance. The framework comprises three main stages: supervised fine-tuning (SFT) with Data Selection, Multiple Chain of Thought (Multi-CoT) tuning, and Direct Preference Optimization (DPO) for de-biasing. We evaluate the performance of the framework through a series of offline experiments on large-scale real-world datasets, as well as online A/B testing. The results indicate significant improvements in both offline and online metrics. Ultimately, the model was deployed in a well-known e-commerce application, yielding substantial commercial benefits.

*Both authors are corresponding authors.

CCS Concepts

• Information systems → Query intent; • Computing methodologies → Natural language processing.

Keywords

Text Classification, E-commerce Retrieval, Search Relevance

ACM Reference Format:

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 Web Conference 2025 (WWW Companion '25)*, April 28-May 2, 2025, Sydney, NSW, Australia. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3701716.3715246>

1 Introduction

E-commerce has become a vital part of daily life, revolutionizing our shopping habits. Platforms like Amazon, Taobao, and JD offer an extensive range of products. The product search system, comprising retrieval and ranking modules, effectively connects user demand with available products[33]. The heart of this system is search relevance, a crucial factor for enhancing user experience, as users expect results that precisely meet their needs[15].

Most previous methods employ BERT-like pre-trained models for search relevance tasks. These models provide a foundational level of context encoding and language understanding, which ensures basic relevance. However, despite their pre-training, these models possess limited world knowledge due to their constrained parameter scale. This limitation hinders their ability to fully comprehend the diverse range of e-commerce products and user expressions. Additionally, their training methods face constraints. Due to the encoder architecture, BERT-like models inherently adopt a discriminative pattern for relevant tasks. This pattern relies on end-to-end data-driven training, making it challenging to optimize the internal reasoning processes[25]. Based on these observations and insights, We choose to base search relevance on large language models (LLMs) instead. Through the pre-training phase, LLMs acquire extensive world knowledge, which is crucial for understanding the wide variety



of products and user queries. Unlike discriminative models, LLMs possess strong reasoning capabilities and employ a generative paradigm to solve application problems. This paradigm allows for a more controllable and optimizable decision-making process regarding the relevance of a product to a user's query, which significantly enhances training flexibility[3].

Due to LLMs' inherent characteristics, applying them to e-commerce search relevance presents three major challenges. (1) Quality Rather Than Quantity Data[13]: LLMs undergo extensive pre-training on large datasets and unify downstream tasks under a generative paradigm. As a result, they already possess an initial capability for text-matching tasks, similar to search relevance tasks, even before specific training. Therefore, when further optimizing for relevant tasks, the quality of the data becomes more critical than its quantity. This necessitates targeted Data Selection from a task perspective. (2) Task-Specific Intermediate Reasoning: As previously mentioned, the generative paradigm allows LLMs to further optimize the decision-making process for downstream tasks. However, this process requires adherence to specific domain task rules. It is challenging to ensure that the model learns and follows these finely-tuned business rules. (3) Optimistic Bias: When applied to relevant tasks, LLMs tend to make more lenient and optimistic judgments, possibly due to their value alignment processes. This characteristic can lead to over-recall, resulting in the mis-exposure of irrelevant products, which negatively impacts user experience[7].

To address these challenges, we propose the LLM-based **RE**levance **FR**amework (**LREF**), which comprises three key components: (1) Data Selection: We perform targeted selection of annotated data by considering the characteristics of relevance tasks, the real-world distribution of online queries, and feedback from the LLM model. Our selection strategy enables the model to achieve better performance with less training data. (2) Multi-CoT Tuning: we guide the LLM to think according to relevant task rules, analyzing complex semantic relationships within query-product pairs and reflecting on incorrect reasoning. This approach positions LLMs as relevance experts, capable of accurately reasoning about user intents.[28] (3) DPO De-biasing: To address the issue of excessive recall due to the LLM's over-optimistic judgments, we model this as a preference optimization problem. By introducing DPO methods, we reinforce its learning of the ambiguous query-product pairs and guide LLMs to make objective and unbiased judgments in borderline cases.[20] We evaluate the performance of LREF with large-scale offline datasets, as well as online A/B testing. The results demonstrate notable enhancements in both offline and online metrics. Ultimately, the model was deployed in a well-known e-commerce application, delivering considerable commercial advantages. The contributions of this paper can be summarized as follows:

- We propose an innovative LLM-based framework for product search relevance, named LREF, which demonstrates advantages over previous BERT-like approaches.
- We introduce a novel and practical Data Selection approach for the LLM-based relevance method.
- We design the Multi-CoT Tuning strategy to optimize the internal reasoning processes for search relevance.

- The DPO strategy is applied to reinforce the LLM's performance in boarding query-product pairs and de-bias LLM's over-recall problem.

In this section, we will discuss the previous studies related to our work: the relevant language models and the supervised fine-tuning of large language models.

1.1 Search Relevance

There is plenty of literature about search relevance technology, and some of the older works focus on Keyword Matching, in their framework, the search engine relies on TF-IDF(Term Frequency-Inverse Document Frequency) [1]. As the development of Natural Language Processing (NLP), the synonyms, and related terms are to be considered by using the Word2Vec[17] and GLoVe[19]. However, these method has the limitation of capturing user intents and analyzing contextual semantics. Over time, Transformer-based models[23] have revolutionized the field, the relevance problems not only solved by language models but also by incorporating user feedback into the learning process[31]. Furthermore, graph-based relevance problem-solving is also presented in improving search relevance.[4]. With the rapid advancement of technology, the model size and complexity are increasing, and approaches like DeBERTa[11] and extended BERT pretraining demonstrate the importance of larger-scale models. However, the constrained knowledge capacity of these approaches limits their ability to fully understand users' intents and complex e-commerce scenarios[32]. Recently, the emergence of LLMs has dramatically expanded model parameters and has a strong performance in natural language tasks[22]. It is essential for industry frameworks to quickly leverage the advantages of the latest models. In our framework, unlike most of the industry end-to-end data fine-tuning[16], we propose a novel LLM-based framework for product search relevance.

1.2 Large Language Model

With the advent of large language models (LLMs), OpenAI's GPT-3[9] introduced few-shot and zero-shot capabilities which enhance model capabilities in domain-specific tasks without requiring extensive task-specific data[10]. However, supervised fine-tuning (SFT) plays a crucial role in achieving optimal performance in targeted applications by adjusting the model's parameters to improve performance for tasks.[18]. Among recent SFT methodologies, Chain-of-Thought (CoT) [28] prompting has been influential in reasoning-intensive tasks. For instance, Self-Consistency CoT [27] improves logical consistency and accuracy in complex problem-solving scenarios. Additionally, the quality of training data has emerged as a decisive factor in the finetuning process[30], the high-quality datasets dramatically enhance the task-solving capabilities of large language models and allow a more effective training process[2, 14]. Moreover, supervised fine-tuning incorporates Reinforcement Learning from Human Feedback (RLHF), enhancing the model's ability to understand user intents[8, 12]. To address the search relevance domain-specific tasks, our three-stage framework includes Data Selection, Chain of Thought (CoT) tuning[28], and direct preference optimization (DPO) [20] for de-biasing. This combined approach allows the models to improve the relevance of domain-specific task ability while maintaining efficiency.

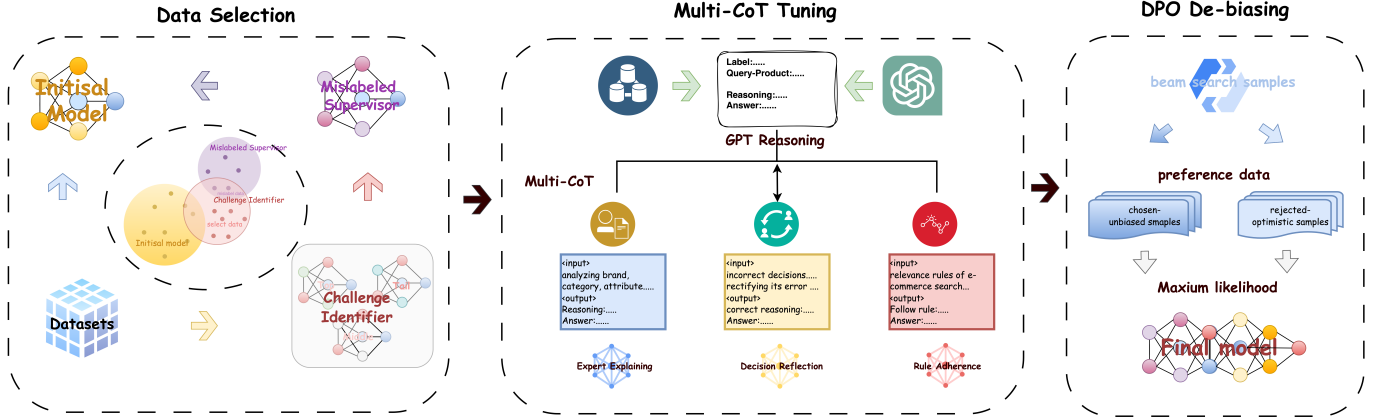


Figure 1: LREF: a novel LLM-based Relevance Framework for Product Search

2 Methods

The task is product search relevance, which can be formulated as a classification problem. Drawing from a real-world e-commerce application, we have developed the ESMTR classification schema, which includes the categories: Exact, Significant, Marginal, Trivial, and Irrelevant. For a given query-product pair, the model is tasked with predicting the appropriate label.

2.1 Methods Overview

In this part, we introduce the proposed LREF method. As shown in Figure 1, LREF is mainly composed of three stages: As LLMs normally require data of high quality, we first adopt a novel and practical Data Selection approach for Supervised fine-tuning (SFT). Subsequently, the Multi-CoT tuning strategy is applied to guide the progress of the LLM reasoning for relevance judgment. Finally, faced with the optimistic bias problem, we tune the model with Direct preference optimization (DPO) for better performance on the marginal cases.

Specifically, the select data module is used to select reliable and influential datasets for the model. The Multi-CoT module bolsters model reasoning ability in search relevance environment and the DPO De-biasing model training model to deal with the boundary problem effectively.

2.2 SFT with Data Selection

Holding initial strong capability for text-matching tasks, LLM is sensitive to data quality[26] rather than relying on vast data, during SFT, we focus on selecting influential data from large human-labeled data. Specifically, we aim to select challenging but not noisy relevant examples. To achieve this, we have specifically designed three auxiliary models: Initial Model, Challenge Identifier, and Mislabel Supervisor. Those three models are initialed from the same LLM, but trained with different data.

The **Initial Model** (IM), built upon an open-source LLM, is fine-tuned using random samples from human-labeled datasets, where common query-product pairs are prevalent. As a result, it performs well on typical examples. In contrast, the **Challenge Identifier** (CI) is designed to recognize more challenging examples and is

trained on a more balanced and diverse set of query-product samples. Specifically, we utilize an e-commerce query online feature to categorize these pairs into three types based on the exploring distribution: top (common), middle, and long tail. We then sample these pairs uniformly to ensure that the middle and long-tail pairs have a significant impact. Specifically, the Challenge Identifier identifies as follows:

$$S_{\text{seed}} = \{x \in D \mid CI(x) \text{ is correct}\} \quad (1)$$

$$S_{\text{challenging}} = \{x \in S_{\text{seed}} \mid IM(x) \neq L(x)\} \quad (2)$$

the Dataset D is the full training dataset, $CI(x)$ is the prediction of the Challenge Identifier, the Labels $L(x)$ is the set of human-labeled classes and the $IM(x)$ denote to the prediction of the Initial Model for x . So firstly we select a seed sample with CI then filter out challenging sample samples with IM.

In the next step, the **Mislabel Supervisor** (MS) is designed to identify and filter out potentially mislabeled and ambiguous annotations within datasets. We randomly select samples from the annotation datasets and employ GPT to ascertain the most confounding classes for these samples. For instance, a typical prompt might be: "In an e-commerce search environment, given the query 'iPhone' and a product described as 'Apple 20W USB-C Phone Charger Original Fast Charger for iPhone,' the current relevance class is 'Trivial.' What would be the most confounding alternative relevance class?" In this context, GPT might suggest 'Marginal.' These ambiguous labels are subsequently used to fine-tune the Mislabel Supervisor, enhancing its ability to predict the most confounding labels. This methodology assists the supervisor in identifying and filtering out annotations that may be erroneously labeled by humans, thereby purifying the training datasets from noise. So the filter out samples where the mislabeled supervisor suggests possible labeling noise. The process is defined as follows:

$$S_{\text{selection}} = \{x \in S_{\text{challenging}} \mid MS(x) \neq L(x)\} \quad (3)$$

In the Data Selection process, we employ these three models to evaluate the entire training dataset. We begin by selecting the samples correctly predicted by the Challenge Identifier as seed samples. From these, we filter out the simple samples that are also correctly

predicted by the Initial Model. Next, we remove potentially mislabeled samples, which are those where the Mislabeled Supervisor's predicted labels match the human annotations. This process yields a set of noise-free, valuable challenging samples. The final selection set S is defined as follows:

$$S_{\text{selection}} = \{x \in D \mid CI(x) \text{ is correct}, IM(x) \neq L(x), MS(x) \neq L(x)\} \quad (4)$$

Finally, we use these samples to further SFT the Initial Model by a Language Model (LM) loss:

$$\mathcal{L}_{lm} = -\sum_{t=1}^T \log P(w_{t+1} \mid w_1, w_2, \dots, w_t) \quad (5)$$

resulting in a model that performs better than one trained on the entire annotated dataset. The specific results of this comparison will be detailed in the experimental section.

2.3 Multi-CoT Tuning

Unlike traditional BERT-like models that primarily depend on end-to-end learning, large language models (LLMs) leveraging the generative paradigm can improve their internal decision-making capabilities for downstream tasks[34]. Nonetheless, achieving this requires strict compliance with domain-specific task guidelines. Ensuring that the model handles the complex e-commerce domain and adheres to these meticulously crafted business rules remains a significant challenge. For product search relevance, we designed the Multiple Chain of Thought Tuning (Multi-CoT Tuning) strategy to address the problem. We introduce three types of CoT: Expert Explaining (EE), Rule Adherence (RA), and Decision Reflection (DR).

Expert Explaining Chain of Thought (EE-CoT) is introduced to offer domain-specific analysis from multiple perspectives. [5] Given a training data point consisting of a user query, product title, and relevance label, represented as (query, title, label), we provide this triplet, including the label, to the GPT model. The model is then tasked with analyzing the query and product from multiple dimensions—such as brand, category, attribute, and keywords—using the relevance label as a guide. This analysis results in an explanation of the EE-CoT. Subsequently, we construct a quadruplet (query, title, label, EE-CoT) and input the (query, title) pair into our LLM, training it to generate the EE-CoT and label. This constitutes a form of thought-guided training.

Furthermore, we developed the **Rule Adherence** Chain-of-Thought (RA-CoT) to guide the LLM in adhering to the relevance rules of a real e-commerce search system. This rule defines query-product relevance through two primary components: product relevance and modifier relevance. Product relevance assesses whether the item meets the basic product type and functional requirements, such as a Samsung S24 being relevant at the product level to a query for "iPhone 15" based on product type and functionality. Modifier relevance focuses on specific descriptive attributes like brand, model, and features; in the previous example, the brands Apple and Samsung do not match, indicating a lack of modifier relevance. Based on the alignment across these two dimensions, a final five-tier relevance judgment is made, categorized as Exact, Significant, Marginal, Trivial, or Irrelevant. We input this complex judgment rule alongside the original sample to form a quadruplet (Rule, query, title, label), which is provided to the GPT model. The

model uses the rule to generate a corresponding reasoning process as the RA-CoT, resulting in a quintuplet (Rule, query, title, label, RA-CoT). Subsequently, we input the (query, title, rule) into the model, training it to generate both the RA-CoT and the label.

In addition, to forward explanation and reasoning, we implemented a **Decision Reflection** Chain of Thought (DR-CoT)[21]. This approach involves leveraging incorrect decisions made by the base LLM to guide the model in rectifying its errors. We provide the GPT model with these incorrect predictions, which are combined with the original sample information to form a tuple (incorrect decision, query, title, label). The model is then tasked with generating the correct reasoning path and identifying the flaws in the initial predictions. This corrective reasoning, referred to as DR-CoT, is subsequently fed back into the base LLM. Specifically, the LLM is later provided with (query, title, incorrect decision) as input, and it is trained to output both the DR-CoT and the correct label.

2.4 DPO De-biasing

After performing Supervised Fine-Tuning (SFT) with Data Selection and Multi-CoT tuning, further analysis of the prediction samples revealed the emergence of an optimistic bias in the LLM model[24]. In relevance tasks, when faced with uncertainty, the LLM tends to classify a marginal query-product pair as significant, thus erroneously deeming it relevant. Specifically, after two stages of training, we found that the model still incorrectly predicted 9% of the training data. Of these incorrect predictions, 70% were misclassified as significant to the query-product pair, resulting in a state of over-recall. However, when we increase the beam size during prediction, we find that 80% of the previously overestimated cases are corrected in the second position of the output. Based on these observations and analyses, we modeled this issue as a preference alignment problem, aiming to shift the LLM's approach from overly optimistic to objective and cautious in relevant tasks. To achieve this adjustment, We employed the Direct Preference Optimization (DPO) method:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}(x, y^+, y^-) [\log \sigma(f_{\theta}(x, y^+) - f_{\theta}(x, y^-))] \quad (6)$$

where $f_{\theta}(x, y)$ is the LLM with parameters θ , (x, y^+, y^-) represents an input and its corresponding preference pair. y^+ is the chosen output and y^- is the rejected one. Specifically, we selected samples from the training data where the LLM, trained by the previous two stages, made incorrect predictions, but the correct answer appeared within the top k positions during the beam search process[29]. We choose the incorrect answer as the rejected data y^- , while the correct answer from beam search is chosen data y^+ . This approach reinforces the model's learning of the preference order among relevant data classes and guides it in making objective and unbiased predictions, particularly in cases of weak relevance.

3 Experiment

This section provides a detailed exploration of the offline and on-line experiments. We first introduce the dataset composition and class distribution of our test datasets. Next, analyze the experiment results with several strong baseline models. Following this, we will show the Data Selection experiences at each selection stage, highlighting their impact. We then compare the different types of CoTs and the DPO improvements of the LLM model. Subsequently, we

Table 1: Offline Evaluation Results

Models	ESMTR classification			Relevant classification		
	Macro F1	Weighted F1	Accuracy	Precision	Recall	F1
BERT	54.09	63.71	63.86	81.21	80.72	80.96
DeBERTa	53.25	64.78	64.78	80.10	85.10	82.52
LLM Base	44.86	59.63	62.80	81.76	78.61	80.16
LREF (DataSelect)	53.00	65.07	65.74	83.68	82.46	83.07
LREF (DataSelect&Multi-CoT)	56.21	66.03	66.23	82.72	86.12	84.39
LREF (DataSelect&Multi-CoT&DPO)	55.90	66.91	67.08	84.12	85.82	84.96

Table 2: Distribution of Test Datasets

Relation	Number	Proportion
Irrelevant	3135	0.95
Trivial	90057	27.29
Marginal	61809	18.73
Significant	135564	41.08
Exact	39435	11.95
total	330000	100

exhibit the online experiences of the model on the JD ab test platform and analyze its effect on user engagement and performance metrics.

3.1 Datasets and Metrics

The test datasets are directly collected from the JD application online users' click logs, with human annotators providing the labels. The datasets consist of 33,000 queries-product pairs and the statistics of the datasets are listed in Table 2. Compared with the online query-product distribution, the test datasets involve a higher proportion of Significant, Marginal, and Trivial classes to evaluate the model's capability in addressing boundary cases. The training datasets are approximately from the same distribution as the test datasets, comprising a total of 5,000,000 query-product pairs.

We evaluate our framework from two perspectives: the ESMTR classification and the irrelevant classification. the ESMTR classifications are divided into five classes and applied for the online product show order. The relevant classification is used to enhance the user experiences in online marketing. We evaluate our framework by weighted F1 scores, macro F1 scores, and accuracy (ACC) metrics. Each evaluation metrics are applied to validate the framework's effectiveness in real-world applications.

3.2 Baselines

For the baselines, we compare our models with several strong baseline models including LLaMA-2-7B [10], BERT [6], and DeBERTa [11]. The evaluation focuses on assessing the fundamental text-processing capabilities of the LLM and determining whether our framework outperforms other benchmarks.

- **BERT**: transformer-based language model designed for understanding the context of words in natural language. We train the model on the full train set.

- **DeBERTa**: Introduced by Microsoft in 2020, it improves upon BERT by addressing limitations in capturing word order and position information. We train the model on the full train set.
- **LLM Base**: We fine-tune the open-source LLM, LLaMA-2 7B, on a full train set and then obtain the LLM Base model.
- **LREF**: the proposed LLM-based Relevance Framework (LREF), initialed from LLaMA-2 7B, and optimized by Data select SFT, Multi-CoT Tuning and DPO de-biasing strategies.

3.3 Implementation Detail

We employed the 7B LLM to build LREF. All training process on 8 H-800 GPUs, with the batch sizes to 16. The warm-up ratio is 0.2 and the deep-speed stage is 1. Furthermore, we use the Adamw optimizer with the learning rate $2e^{-5}$. The max length of the model training is 500 and the epoch is 8. To overcome the overfitting, we save the best checkpoints in the end and use validation datasets as a split of 10 percent of the training datasets. As for Dpo training the α is set to 0.65. The epoch is to be set as 2 to prevent overfitting. For the inference part, we utilize vLLM with the following settings with the temperature set as 0 for deterministic outputs.

3.4 Offline Evaluation

3.4.1 Overall Performance. As shown in Table 1, our proposed LREF method achieves significant improvements compared to the baselines. Furthermore, taking LLM Base as a reference point, we observe that both BERT and DeBERTa outperform it. This indicates that the straightforward application of LLMs to relevant tasks, specifically by fine-tuning LLMs with all annotated data (SFT), does not inherently surpass the performance of classical BERT-like discriminative models. Only after we address the need for high-quality data in SFT, optimize the internal reasoning steps, and mitigate optimistic bias, our proposed LREF, which also leverages LLMs, achieve optimal performance.

3.4.2 Effectiveness of Data Selection in SFT. In Table 1, LREF (DataSelect) is the model that SFT with of Data Selection method. Specifically, through the proposed Data Selection method, we obtain about 0.5m noise-free, valuable challenging samples from the train set (about 10% selected), and sequentially fine-tune the Initial Model, which is already trained on a random 150,000 samples. As shown in Table 1, the comparison between LLM Base and LREF(DataSelect) demonstrates the effectiveness of our Data Selection method for LLM applying to relevant tasks.

Table 3: The Ablation Study of Data Selection SFT

Models	ESMTR classification			Relevant classification		
	Macro F1	Weighted F1	Accuracy	Precision	Recall	F1
Full Data	44.86	59.63	62.80	81.76	78.61	80.16
DataSelect (IM)	46.29	59.89	60.10	78.74	83.09	80.85
DataSelect (IM&CI)	52.81	63.91	63.88	83.34	82.27	82.81
DataSelect (IM&CI&MS)	53.00	65.07	65.74	83.68	82.46	83.07

Table 4: The Ablation Study of Multi-CoT Tuning

Models	ESMTR classification			Relevant classification		
	Macro F1	Weighted F1	Accuracy	Precision	Recall	F1
w/o Multi-CoT	53.00	65.07	65.74	83.68	82.46	83.07
Multi-CoT (EE)	54.30	64.58	64.38	83.02	85.01	84.00
Multi-CoT (EE&RA)	55.40	65.27	65.25	83.27	84.24	83.75
Multi-CoT (EE&RA&DR)	56.21	66.03	66.23	82.72	86.12	84.39

Table 5: The Analysis for DPO-biasing

Class	LLM Base			LREF			BERTmodel			DeBERTamodel		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Marginal	66.88	50.13	57.31	69.02	64.12	66.48	62.86	62.30	62.58	63.88	63.73	63.80
Significant	61.39	73.52	66.91	66.91	71.18	68.98	64.26	64.63	64.45	63.82	66.94	65.34

Table 3 shows the further ablation study of Data Selection in SFT. DataSelect(IM) is the Initial Model(IM) which is built on the same open-source LLM with LLM Base, and trained on random 0.15m samples. DataSelect(IM&CI) is the model trained from the Initial Model on 0.56m challenging but noisy samples selected by CI and IM. DataSelect(IM&CI&MS) represents the model trained from the Initial Model on 0.5m challenging and noise-free samples selected by IM, CI, and MS. The results in Table 3 show LLM Base, the LLM trained on the full train set, only matches the performance of the Initial Model trained on a randomly selected subset of data. However, with the introduction of CI and MS, the data is progressively refined, and the performance of LLM SFT is gradually improved. Ultimately, in the SFT stage, LREF significantly surpasses the performance of the LLM trained on the full dataset by using only a portion of high-quality data.

3.4.3 Effectiveness of Multi-CoT Tuning. As shown in Table 1, LREF (DataSelect&Multi-CoT) reports the performance of Multi-CoT Tuning after SFT with Data Selection. It outperforms LREF (DataSelect) significantly, which proves the effectiveness of Multi-CoT tuning which focuses on optimizing the internal reasoning progress of LLM for relevant tasks.

Table 3 shows the further ablation study of Multi-CoT Tuning. In this table, w/o Multi-CoT means LLM fine-tuned by the selected data, the first stage of LREF. The Multi-CoT (EE) is the LLM trained subsequently from the first stage on the Expert Explaining (EE) Chain of Thought (CoT) data. Multi-CoT (EE&RA) and Multi-CoT (EE&RA&DR) indicate the successive incorporation of Rule Adherence (RA) CoT and Decision Reflection (DR) CoT into the training process, respectively. With the guidance of Expert Explaining, the

LLM became more adept at understanding e-commerce search scenarios, leading to significant improvements in the Relevant classification metric, which directly impacts user experience. Furthermore, the incorporation of Rule Adherence enabled the LLM to follow pre-defined standards during the reasoning process, enhancing overall performance. Finally, with Decision Reflection, the model engaged in targeted reflection on erroneous examples, elevating the effectiveness of the Multi-CoT strategy to a notably higher level.

3.4.4 Effectiveness of DPO De-biasing. Through the experiments, we observed that when applied to relevance tasks, LLMs tend to make lenient and optimistic judgments. Specifically, compared to BERT-like baselines, LLMs are more prone to misclassifying Marginal cases as Significant, adopting a more optimistic stance in evaluating the relevance between products and queries. As shown in Table 5, the LLM Base exhibits lower recall for the Marginal class and higher precision for the Significant class. Consequently, our LREF approach employs DPO to mitigate this optimistic bias. As indicated by the data in the table 1, incorporating DPO-debiasing, in addition to Data Selection SFT and Multi-CoT Tuning, further enhances the overall performance of our proposed LREF. Moreover, when examining the precision and recall across different categories, we found significant improvements in Marginal recall and Significant precision. This indicates that the corresponding relevance methods can better control the mis-exposure of non-significantly relevant samples.

3.5 Online Evaluation

3.5.1 Online Deployment. To reduce the response latency of online deployment, we leverage the data distillation method for

transferring LLM ability in online deployment, we employ LLM to annotate approximately 2 billion datasets and then transfer the knowledge to a BERT-based model for efficient online serving. Additionally, the relevance classification is sent to the rank module, which organizes product orders within each relevance class. This means that the relevance prediction results will influence the product display order in the JD apps. For these irrelevance products it will be deleted in the ranking stage and not be displayed under the query search pages. This ensures that irrelevant items are filtered out in the ranking pipeline, improving the user experience and online efficiency.

3.5.2 Online Performance. To evaluate our framework, we deployed it on the JD A/B test platform. We randomly select a 20% traffic group as the test group to employ our LREF approach. Another 20% served as the base group which utilized the previous BERT-based model. For the fair comprising, we continuously monitored experimental performance on the platform. To avoid the effects of traffic fluctuations, we set at least a 7-day testing period. For online evaluation, the following business metrics were used:

$$\begin{aligned} UVvalue &= \frac{GMV}{UV}, \\ UCVR &= \frac{Orderlines}{UV}, \\ UCTR &= \frac{Clicks}{UV}, \end{aligned} \quad (7)$$

To be specific, the UCVR is the conversion rate of users and UCTR is the click rate of users. Furthermore, Relevance Satisfaction is also used as a key metric, it measures how the displayed products match user intent based on the query. Expert annotators measure the relevance satisfaction, they determine the relevance satisfaction by reviewing each query and evaluating the displayed products case by case. Around 20,000 cases are annotated to generate relevance satisfaction in test and base groups. we observe that Our LREF

Table 6: Online improvements of the LREF.

Models	UV value	UCVR	UCTR	RS
Online	-	-	-	-
LREF	+0.023%	+0.209%	+0.120%	+1.016%

frame shows demonstrates great improvements in both the UCVR and Relevance Satisfaction. (1) The great improvement in Relevance Satisfaction shows that more relevant products are displayed in the apps, enhancing the user experiences. (2)The UCTR improvement indicates that users are provided with more relevant products that match their intents leading to higher satisfaction click rates (3) The growth of UCVR indicates that presenting more relevant products leads to an increase in user clicks and purchases. These results underline that employing the LREF approach improves both user satisfaction and business efficiency.

4 Conclusion

This paper introduces a novel framework, which is called LLM-based **Relevance Framework (LREF)**, designed to enhance search relevance in e-commerce applications. Aiming to address the challenges of applying LLMs to query-product relevance classification

tasks, LREF offers a novel way to select high-quality data in large noisy human annotation datasets, involves an internal guide for task-specific reasoning, and applies a DPO-based De-biasing approach to mitigate LLM’s optimistic bias in relevance decisions. We conduct a series of offline experiments on large-scale real-world datasets and online A/B testing. The experiments show that our framework achieves significant improvements in both offline and online metrics. Finally, the model is deployed on a well-known e-commerce application and achieves significant results.

References

- [1] Akiko Aizawa. 2003. An information-theoretic perspective of tf-idf measures. *Information Processing & Management* 39, 1 (2003), 45–65.
- [2] Yihan Cao, Yanbin Kang, Chi Wang, and Lichao Sun. 2023. Instruction Mining: Instruction Data Selection for Tuning Large Language Models. *arXiv preprint arXiv:2307.06290* (2023).
- [3] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology* 15, 3 (2024), 1–45.
- [4] Nurendra Choudhary, Edward W Huang, Karthik Subbian, and Chandan K Reddy. 2024. An interpretable ensemble of graph and language models for improving search relevance in e-commerce. In *Companion Proceedings of the ACM on Web Conference 2024*. 206–215.
- [5] Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. A survey of chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402* (2023).
- [6] Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [7] Eva Eigner and Thorsten Händler. 2024. Determinants of llm-assisted decision-making. *arXiv preprint arXiv:2402.17385* (2024).
- [8] Kavin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306* (2024).
- [9] Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines* 30 (2020), 681–694.
- [10] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010* (2023).
- [11] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654* (2020).
- [12] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2023. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925* (2023).
- [13] Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2023. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. *arXiv preprint arXiv:2308.12032* (2023).
- [14] Ziche Liu, Rui Ke, Feng Jiang, and Haizhou Li. 2024. Take the essence and discard the dross: A Rethinking on Data Selection for Fine-Tuning Large Language Models. *arXiv preprint arXiv:2406.14115* (2024).
- [15] Ziyang Liu, Chaokun Wang, Hao Feng, Lingfei Wu, and Liqun Yang. 2022. Knowledge distillation based contextual relevance matching for e-commerce product search. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*. 63–76.
- [16] Navid Mehrdad, Hrushikesh Mohapatra, Mossaab Bagdouri, Prijith Chandran, Alessandro Magnani, Xunfan Cai, Ajit Puthenpuussery, Sachin Yadav, Tony Lee, ChengXiang Zhai, et al. 2024. Large Language Models for Relevance Judgment in Product Search. *arXiv preprint arXiv:2406.00247* (2024).

- [17] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26 (2013).
- [18] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435* (2023).
- [19] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.
- [20] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems* 36 (2024).
- [21] Matthew Renze and Erhan Guven. 2024. Self-Reflection in LLM Agents: Effects on Problem-Solving Performance. *arXiv preprint arXiv:2405.06682* (2024).
- [22] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [23] A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).
- [24] Adam R Villafior, Zhe Huang, Swapnil Pande, John M Dolan, and Jeff Schneider. 2022. Addressing optimism bias in sequence modeling for reinforcement learning. In *international conference on machine learning*. PMLR, 22270–22283.
- [25] Jiajia Wang, Jimmy Xiangji Huang, Xinhui Tu, Junmei Wang, Angela Jennifer Huang, Md Tahmid Rahman Laskar, and Amran Bhuiyan. 2024. Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges. *Comput. Surveys* 56, 7 (2024), 1–33.
- [26] Jiahao Wang, Bolin Zhang, Qianlong Du, Jiajun Zhang, and Dianhui Chu. 2024. A Survey on Data Selection for LLM Instruction Tuning. *arXiv preprint arXiv:2402.05123* (2024).
- [27] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171* (2022).
- [28] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.
- [29] Sam Wiseman and Alexander M Rush. 2016. Sequence-to-sequence learning as beam-search optimization. *arXiv preprint arXiv:1606.02960* (2016).
- [30] Mengzhou Xia, Sathika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333* (2024).
- [31] Su Yan, Wei Lin, Tianshu Wu, Daorui Xiao, Xu Zheng, Bo Wu, and Kaipeng Liu. 2018. Beyond keywords and relevance: a personalized ad retrieval framework in e-commerce sponsored search. In *Proceedings of the 2018 World Wide Web Conference*. 1919–1928.
- [32] Chunyuan Yuan, Ming Pang, Zheng Fang, Xue Jiang, Changping Peng, and Zhangang Lin. 2024. A Semi-supervised Multi-channel Graph Convolutional Network for Query Classification in E-commerce. In *Companion Proceedings of the ACM on Web Conference 2024*. 56–64.
- [33] 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*. 416–420.
- [34] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493* (2022).