LLM-based Keyphrase-augmented Framework for Semantic Relevance Assessment in E-commerce

Guoliang Zhang⋆[1], Gang Zhao*[2], Zhiyuan Zeng[2], Songyan Liu[2], Haoyue Zhang[1], Hui Zhao[2], Tianshu Wu[2], PengjieWang[2], Jian Xu[2], Bo Zheng[2]🖂, and Baolin Liu[1]🖂

[1] University of Science and Technology Beijing, China
[2] Alibaba Group, China
{zhangguoliang,zhanghaoyue}@xs.ustb.edu.cn, liubaolin@ustb.edu.cn
{zilong.zg,zengzhiyuan.zzy,moxuan.lsy,shuqian.zh}@alibaba-inc.com
{shuke.wts,pengjie.wpj,xiyu.xj,bozheng}@alibaba-inc.com

Abstract. Relevance assessment in search advertising is essential for enhancing user experience. However, current relevance assessment methods often perform sub-optimally in the face of redundant information in the ad titles. This issue can be largely ascribed to the problem of encoding imbalance, in which the redundant information is overvalued. To address these challenges, we propose a novel **K**eyphrase-**E**nhanced **S**emantic **R**elevance (**KESR**) framework. The key insight of KESR is to refine and enhance the ad title semantics by offline keyphrase extraction, thereby improving the performance of the online relevance assessment model. The KESR framework comprises two primary components: an offline keyphrase extraction module and an online keyphrase-enhanced relevance module. To achieve superior performance in keyphrase extraction, particularly for long-tail samples, we propose a LLM-based keyphrase extraction module (**LLM-KPE**). Furthermore, to address the discrepancy between human prior keyphrase criteria and downstream relevance task requirements for keyphrases, inspired by Reinforcement Learning from Human Feedback (RLHF), we propose a **R**einforcement **L**earning from **D**ownstream **F**eedback (**RLDF**) method. By aligning the LLM-generated keyphrases with those required for downstream relevance tasks, the LLM can generate keyphrases better suited for relevance tasks. Experimental results demonstrate the effectiveness of KESR in keyphrase extraction, relevance assessment, and online deployment.

Keywords: semantic matching · keyphrase extraction · large language model · preference alignment · e-commerce

1 Introduction

As one of the largest e-commerce platforms globally, Taobao provides online shopping services to millions of users. In e-commerce search engines like Taobao,

⋆ Guoliang Zhang and Gang Zhao contributed equally to this work.
🖂 Baolin Liu and Bo Zheng are the corresponding authors.

precisely targeted advertisements can not only enhance user engagement but also improve advertisers' return on investment [4]. To optimize advertising performance, accurately assessing relevance has become a crucial research focus. Consequently, numerous methods have been proposed to achieve more accurate relevance assessments [5, 15].

In relevance assessment tasks, classical neural network methods can be broadly divided into two categories: Interaction-based models [7] and Representation-based models [6]. Interaction-based models take the concatenation of queries and titles as input to perform full interaction, which effectively captures relevance features. However, they are challenging to deploy online due to the inability to pre-store embeddings, resulting in significant latency issues. In contrast, Representation-based models independently encode queries and titles into embeddings, assessing semantic similarity using lightweight operations, such as cosine similarity. This paradigm allows for efficient real-time computing, as title embeddings can be precomputed.

The existing methods following the Representation-based models primarily focus on alleviating the deficiency of insufficient feature cross-interaction capability. For instance, to circumvent the limitations of feature crossing, [13] proposed distilling knowledge from a teacher model into the student model. To further enhance cross-interaction capabilities, [14] employed a triple-encoder structure and combined feature screening, and engineering optimization techniques. Though promising, these methods still struggle to handle redundant information[3]. This struggle is primarily caused by encoding imbalance, in which redundant features are overvalued. Therefore, refining the semantic content of ad titles is crucial.

In this paper, we propose a **K**eyphrase-**E**nhanced **S**emantic **R**elevance (KESR) framework to address the aforementioned challenges. The key insight of KESR is to refine and enhance the semantics of ad titles by extracting keyphrases that are highly important to users, thereby improving the effectiveness of relevance assessment. Specifically, KESR mainly consists of two modules: an offline ad title keyphrase extraction module and an online keyphrase-enhanced relevance module. For keyphrase extraction, we collect user click behavior data to construct a keyphrase extraction dataset. Considering the lack of labeled data for the long-tail samples in the e-commerce scene, we propose to model the keyphrase extraction task using a LLM containing rich world knowledge. Additionally, we adopt the output word order of LLM to model the importance of keyphrases, thus learning the relative importance among the keyphrases. For the keyphrase-enhanced relevance model, we propose a triple-encoder architecture consisting of query, ad title, keyphrase encoders, and adaptive gating for relevance score fusion. The keyphrase encoder is used to refine semantic information and filter out irrelevant information. The ad title encoder makes a semantic complement to avoid ignoring the potential information in the original title.

Furthermore, to address the discrepancy between the keyphrase sequences generated based on predetermined thresholds and the optimal keyphrase se-

[3] We define redundant information as follows: given a query set {{a}, {b}} and a title {a, b, c}, the term {c} is considered redundant for the search advertising system.

quences required by downstream tasks, we propose a novel **R**einforcement **L**earning from **D**ownstream **F**eedback (RLDF) method. Specifically, the current keyphrase threshold is set based only on statistical coverage and is not suitable for every specific sample. Inspired by Reinforcement Learning from Human Feedback (RLHF), we propose to employ RLDF to adjust for this discrepancy. Concretely, we leverage the trained keyphrase-enhanced relevance model as the reward function and design feedback scores for different criteria keyphrase sequences. Then, we calibrate the LLM by fitting its predicted probability rankings for different criteria keyphrase sequences to the corresponding keyphrase sequence score rankings, thereby aligning the keyphrases generated by LLM with the preference of the downstream relevance assessment model.

The main contributions of our work are summarized as follows:

- To address the issue of diminished performance in relevance assessment models caused by redundant ad titles, we propose a Keyphrase-Enhanced Semantic Relevance (KESR) framework, aiming to augment ad title semantics and improve relevance assessment performance.
- To address the discrepancy between keyphrases derived from human prior criteria and those necessary for downstream relevance tasks, we propose a RLDF method that optimizes the LLM to generate keyphrases that are better suited to downstream relevance tasks.
- Extensive experiments on keyphrase extraction, relevance assessment, and online deployment demonstrate the effectiveness of our KESR framework.

2 Related Work

Keyphrase Extraction. BERTTagKPE combine BERT's character-level encoding with CRF's word-level information integration and frames keyphrase recognition as a BIO sequence labeling task. KIEMP [12] identifies candidate phrases and ranks them to determine the final set of keyphrases. Boudin et al. [1] demonstrate that keyphrases can facilitate BM25 retriever. However, current methods have some limitations in the application of e-commerce scenarios, especially in the lack of adaptation to long-tail samples.

Text Matching. DSSM [5] uses separate deep fully connected networks to independently encode queries and documents into embeddings, assessing similarity through cosine similarity. Sentence-BERT [39] adopts a shared BERT model, to encode queries and titles separately. However, existing methods struggle to deal with the redundancy information in the e-commerce scene.

Preference Alignment. These preference alignment methods optimize the compatibility of neural network reward functions with preference datasets based on reward models [2], and then fine-tuneLLM to maximize the given reward using reinforcement learning (PPO [10]) or more stable supervised learning (PRO [11]). Most existing methods aim at aligning human preferences, but do not consider the deviation of downstream tasks from upstream.

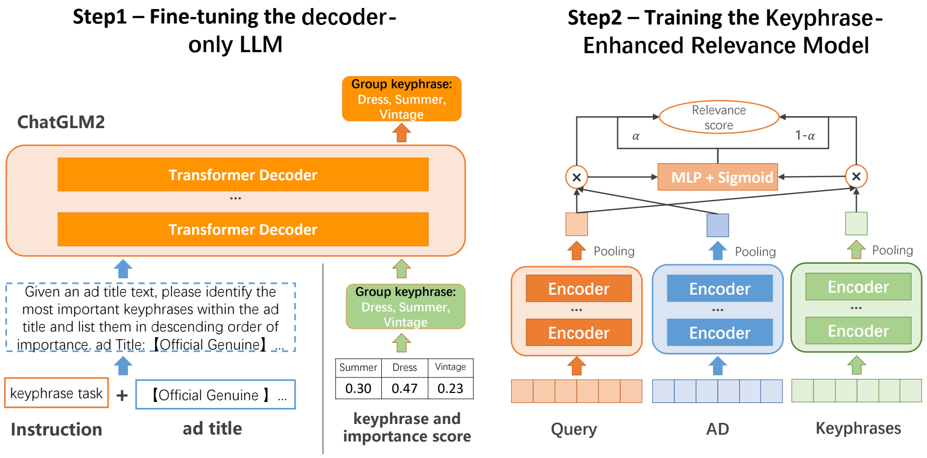

Step1 – Fine-tuning the decoder-only LLM
ChatGLM2
Group keyphrase: Dress, Summer, Vintage
Transformer Decoder
...
Transformer Decoder
Given an ad title text, please identify the most important keyphrases within the ad title and list them in descending order of importance. ad Title: 【Official Genuine】 ...
keyphrase task
+
【Official Genuine 】 ...
Instruction
ad title
Group keyphrase: Dress, Summer, Vintage
Summer
Dress
Vintage
0.30
0.47
0.23
keyphrase and importance score
Step2 – Training the Keyphrase-Enhanced Relevance Model
Relevance score
α
1-α
MLP + Sigmoid
×
×
Pooling
Pooling
Pooling
Encoder
...
Encoder
Encoder
...
Encoder
Encoder
...
Encoder
Query
AD
Keyphrases

Fig. 1: The architecture of KESR consists of the LLM-based Keyphrase Extraction Module (LLM-KPE) and Keyphrase-Enhanced Relevance Module (KERM).

3 Methodology

The overall framework is shown in Figure 1. Let $f_\theta(\mathbf{q}, \mathbf{t})$ denote the previous relevance assessment model, where θ indicates the trainable parameters. The assessment model optimizes θ with the training data $\mathcal{A}(\mathbf{q}, \mathbf{t})$ in a comparative learning way. We construct the keyphrase $\mathcal{C}(\mathbf{t}, \mathbf{k})$ from $\mathcal{A}$ using a regularized method and leverage $\mathcal{C}$ to fine-tune the LLM $g_\theta(\mathbf{t})$. After incorporating the LLM-enhanced keyphrases $\mathcal{C}'(\mathbf{t}, \mathbf{k})$, the new training data becomes $\mathcal{B}(\mathbf{q}, \mathbf{t}, \mathbf{k})$. We utilize this dataset and train a new assessment model $f'_\theta(\mathbf{q}, \mathbf{k}, \mathbf{t})$ to assess the relevance between queries and keyphrase-enhanced ad titles.

$$g_\theta : Train(\mathcal{C}) => f'_\theta : Train(\mathcal{B}) \tag{1}$$

Finally, we use this f'_θ similarity score to further tune the g_θ to accommodate the user click feedback in keyphrase extraction.

3.1 LLM-based Keyphrase Extraction for Relevance Learning

Data Construction. We aim to capture user behavior trends through keyphrases in ad titles. We gather historical search queries focused on ad titles, choosing query-title pairs with at least 300 cumulative clicks per title. Using an e-commerce tokenization model from a public aliNLP platform, we break down queries into semantic units as potential keyphrases. We then filter these, merging, deduplicating, removing stop words, and refining with regular expressions to exclude phrases not in the corresponding ad title group. Finally, we compile the filtered phrases with their related ad titles for each group.

To mitigate bias from ad title exposure, we rank group phrases by frequency and select the top phrases as keyphrase tags such that their cumulative frequency reaches 75%. We define the frequency f_k for the phrase as $f_k = \frac{n_k}{\sum_{k' \in K} n_{k'}}$. The n_k is the count of a phrase's occurrence in search click queries, and $\sum_{k' \in K} n_{k'}$ is the sum of occurrences of all valid phrases in the ad title. The 75% frequency coverage threshold is determined based on case analysis.

We model importance through frequency $\mathcal{I}_{k_i} = \frac{e^{f_{k_i}}}{\sum_{j=1}^{n} e^{f_{k_j}}}$. Here, we obtain $\mathcal{C}(\mathbf{t}, \mathbf{k})$, where $\mathbf{k} = k_1, k_2, ..., k_n$ in descending order of importance score.

Supervised Fine-tuning of LLM. We fine-tuned the LLM using $\mathcal{C}(\mathbf{t}, \mathbf{k})$ to enhance its performance in keyphrase extraction within e-commerce contexts. To learn the relative importance among the keyphrases, we model LLM to rank the generated keyphrases $\mathbf{k}$ according to their importance. We use the prompt, as shown in Figure 1, to guide LLM in performing keyphrase extraction tasks. We concatenate the prompt $\mathbf{p}$ and the ad title $\mathbf{t}$ as the model input. The cross-entropy loss function is optimized as follows:

$$\mathcal{L}_{LLM} = -\frac{1}{N} \sum_{i=1}^{N} \log P_{\theta_{\mathrm{LLM}}}(k_i | \mathbf{p}, \mathbf{t}, k_{j<i}) \tag{2}$$

Keyphrase Generation Inference. In the inference stage, we set the hyper-parameters `num_beams` $= 1$ and `temperature` $= 1.0$ to perform greedy search:

$$\mathbf{k} = \arg\max \mathrm{LLM}(k_i \mid k_{j<i}, \mathbf{p}, \mathbf{t}) \tag{3}$$

We leverage LLM to generate pseudo-labels $\mathcal{C}'(\mathbf{t}, \mathbf{k})$ for ad titles with sparse query data in $\mathcal{A}(\mathbf{q}, \mathbf{t})$, resulting in the augmented dataset $\mathcal{B}(\mathbf{q}, \mathbf{t}, \mathbf{k})$.

3.2 Keyphrase-Enhanced Relevance Model

Encoder Design. We utilize Transformer-based encoders (E_q, E_t, E_k) to convert query, ad title, and keyphrases into embedding representations. The transformation process is described by the following formulas:

$$z_q, z_t, z_k = \mathrm{Pooling}(E_q(\mathbf{q})), \mathrm{Pooling}(E_t(\mathbf{t})), \mathrm{Pooling}(E_k(\mathbf{k})) \tag{4}$$

Where $\mathbf{q}, \mathbf{t}, \mathbf{k}$ represent the query, ad title, and keyphrases, respectively. Our default pooling strategy is CLS.

Weakly Supervised Contrastive Training. By integrating the keyphrases generated by LLM into $\mathcal{A}(\mathbf{q}, \mathbf{t})$, we can obtain $\mathcal{B}(\mathbf{q}, \mathbf{t}, \mathbf{k})$. Our model adopts a contrastive learning approach to take advantage of the plentiful unsupervised data for representation learning. The training objective for $(\mathbf{q_i}, \mathbf{t_i^+}, \mathbf{k_i^+})$ with a mini-batch of N pairs is:

$$\mathcal{L}_i = -log \frac{e^{sim(\mathbf{q_i}, \mathbf{t_i^+}, \mathbf{k_i^+})/\tau}}{\sum_{j=1}^{N} e^{sim(\mathbf{q_i}, \mathbf{t_j^+}, \mathbf{k_j^+})/\tau}} \tag{5}$$

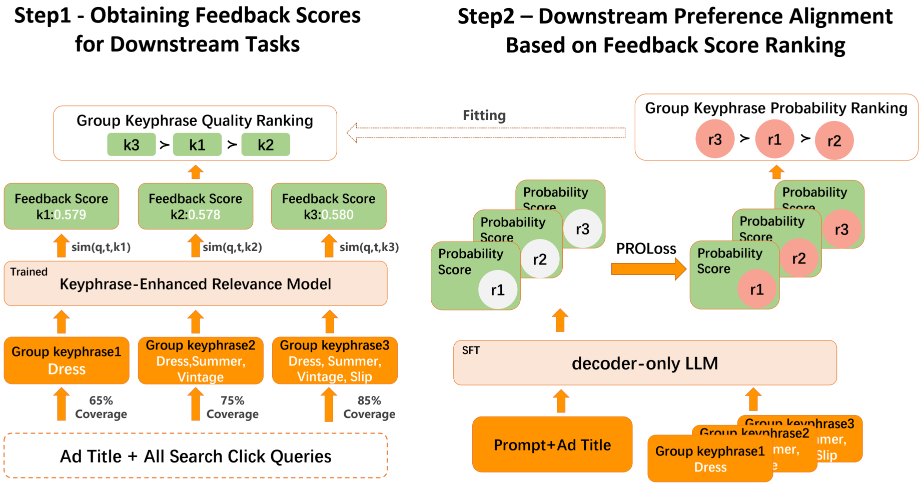

Step1 - Obtaining Feedback Scores for Downstream Tasks
Step2 – Downstream Preference Alignment Based on Feedback Score Ranking
Group Keyphrase Quality Ranking
k3 ≻ k1 ≻ k2
Fitting
Group Keyphrase Probability Ranking
r3 ≻ r1 ≻ r2
Feedback Score k1:0.579
Feedback Score k2:0.578
Feedback Score k3:0.580
sim(q,t,k1)
sim(q,t,k2)
sim(q,t,k3)
Trained
Keyphrase-Enhanced Relevance Model
Group keyphrase1 Dress
Group keyphrase2 Dress,Summer, Vintage
Group keyphrase3 Dress, Summer, Vintage, Slip
65% Coverage
75% Coverage
85% Coverage
Ad Title + All Search Click Queries
Probability Score
r1
r2
r3
PROLoss
SFT
decoder-only LLM
Prompt+Ad Title

Fig. 2: Illustration of the process of Reinforcement Learning from Downstream Feedback (RLDF).

Here, $\mathbf{t_i^+}, \mathbf{k_i^+}$ is the positive sample for $\mathbf{q_i}$, and τ is the temperature parameter. We take a cross-entropy objective with in-batch negatives.

Relevance Calculation. We calculate the query-ad title relevance score (s_{q2t}) and query-keyphrases relevance score (s_{q2k}) respectively, and take the weighted sum of the two through the adaptive weights α as the final result:

$$sim(\mathbf{q}, \mathbf{t}, \mathbf{k}) = \alpha \cdot s_{q2t} + (1 - \alpha) \cdot s_{q2k} \tag{6}$$

In this formula, $s_{q2t} = \frac{z_q \cdot z_t}{\|z_q\| \|z_t\|}$ and $s_{q2k} = \frac{z_q \cdot z_k}{\|z_q\| \|z_k\|}$ are cosine similarity measures, while z_q, z_t, z_k are the embeddings for the query, ad title, and keyphrases, respectively. The α, which dictates the relative contribution of each relevance score, is computed as:

$$\alpha = \sigma(W_g \cdot [s_{q2t} \oplus s_{q2k}]) \tag{7}$$

Where $\oplus$ indicates concatenation operation, σ represents the sigmoid activation function ensuring α is within the $(0, 1)$ range, and W_g is the trainable parameters.

3.3 Preference Alignment Based on Downstream Performance

There exists a discrepancy between the keyphrase sequences generated based on predetermined thresholds and the optimal keyphrase sequences required by downstream tasks. Specific challenges include: 1)The current keyphrase threshold is only set based on statistical coverage and empirical judgment. 2)A uniform threshold not be suitable for every specific relevance assessment.

Previous research treats this as a hyperparameter problem, focusing on selecting the optimal fixed N of predicted keyphrases. In this study, we propose RLDF to adaptively adjust for this discrepancy. We collect downstream feedback (**DF**) scores for keyphrase sequences with different coverage thresholds and obtain quality rankings. Subsequently, we use LLM's predicted probability ranking to fit the quality ranking and learn downstream preferences.

Feedback Score Design. The quality of keyphrases is difficult to assess manually. To address this, we design the feedback score formula to model the feedback signal. Specifically, we calculate the feedback score through three steps:(1) Select $\{\mathbf{q}\}$,$\mathbf{t}, \mathbf{k}^j$ at a certain coverage rate;(2) For each $\mathbf{q}$, compute the relevance score $sim(\mathbf{q}^i, \mathbf{t}, \mathbf{k}^j)$; (3) Aggregate and average all relevance scores to calculate the feedback score for $\mathbf{k}^j$.

$$Score(\mathbf{k}^j) = \frac{1}{N}\sum_{i=1}^{N} sim(\mathbf{q}^i, \mathbf{t}, \mathbf{k}^j) \tag{8}$$

Implementation of Preference Alignment Method. Using the feedback score formula, we can calculate feedback scores for keyphrase sequences at different coverage rates in the ad title. Subsequently, we rank these sequences according to their DF scores, obtaining a downstream keyphrase sequence preference ranking $\mathbf{k}^1 \succ \mathbf{k}^2 \succ \cdots \succ \mathbf{k}^n$, which represents the quality ranking. Specifically, we obtain a downstream feedback dataset $\mathcal{D}(\mathbf{t}, \mathbf{k}^1 \cdots \mathbf{k}^n, \mathbf{score}^1 \cdots \mathbf{score}^n)$.

We use the PRO algorithm [11] to optimize the trained LLM, aligning the probability ranking of different coverage keyphrase sequences generated by the LLM with the downstream quality ranking. We define $r_{\pi_{PRO}}(\mathbf{t}, \mathbf{k}^j)$ as a function parameterized by our desired LLM$_{\pi_{PRO}}$, representing the probability recognition score for a certain keyphrase sequence:

$$r_{\pi_{PRO}}(\mathbf{t}, \mathbf{k}^j) = \frac{1}{|\mathbf{k}^j|}\sum\nolimits_{i=1}^{|\mathbf{k}^j|} logP_{\pi_{PRO}}(k_i^j|\mathbf{t}, k_{<i}^j) \tag{9}$$

Here, LLM$_{\pi_{PRO}}$ calculates this keyphrase sequence response score by multiplying the probability of each token generated by π_{PRO}. We define the following model objective to enhance the LLM to generate high-quality keyphrase sequences:

$$\begin{aligned} P(\mathbf{k}^{1,\cdots,n}|t) &= \prod_{j=1}^{n-1} P(\mathbf{k}^{j,j+1:n}|\mathbf{t}) \\ &= \prod_{j=1}^{n-1} \frac{exp(r(\mathbf{t}, \mathbf{k}^j))}{\sum_{i=j}^{n} exp(r(\mathbf{t}, \mathbf{k}^i))} \end{aligned} \tag{10}$$

In this formula, $\mathbf{k}^1$ is initially considered as a positive response, while the rest are negative responses. Then, discarding $\mathbf{k}^1$, $\mathbf{k}^2$ is considered positive, with the remainder as negatives. This process is repeated until no responses remain.

Table 1: Statistics of the constructed dataset. The key-term refers to the number of unique phrases. The "click-train, rele-test, click-train, RLDF-train, key-test" represents the click behaviors, relevance assessment test, ad-keyphrase train, keyphrase preference alignment train, and ad-keyphrase test, respectively.

Dataset	click-train	rele-test	key-train	RLDF-train	key-test
Sample	179M	5M	35K	33K	2K
query	5M	107K	-	-	
title	21M	559K	35K	33K	2K
key-term	-	93K	19K	15K	2K

Optimization. With the above probability $P(\mathbf{k}^{j,j+1:n}|\mathbf{t})$, we then compute the corresponding cross-entropy loss $\ell(t,\mathbf{k}^{j,j+1:n})$. We aggregate these losses to obtain the final PRO loss:

$$\mathcal{L}_{PRO} = -\sum\nolimits_{j=1}^{n-1} log \frac{exp(r_{\pi_{PRO}}(\mathbf{t},\mathbf{k}^{j}))}{\sum\nolimits_{i=j}^{n} exp(r_{\pi_{PRO}}(\mathbf{t},\mathbf{k}^{i}))} \tag{11}$$

We optimize this loss function to guide the LLM in generating the optimal keyphrase sequences required for the downstream relevance task.

4 Experiment

4.1 Experiment Setup

Dataset. As shown in Table 1, we collect user click logs from online search advertising systems to construct the click-train datasets. The rele-test contains 5 million query-item pairs, labeled Good (relevant) or Bad (irrelevant) by experienced human annotators. The key-train and RLDF-train datasets are based on the previously mentioned method. The key-test is obtained by sampling data with click counts greater than 600 from key-train. We guarantee that there is no overlap between train datasets and test datasets.

Evaluation Metrics. For relevance assessment, we use the Area Under ROC Curve (AUC) as the evaluation metric. For Keyphrase extraction, we use the Precision, Recall, and F1 metrics. Online performance is assessed using business metrics, including Click Through Rate (CTR) and click Value Rate (CVR).

Implementation Details. In our relevance model experiment, each encoder uses a six-layer Transformer Encoder. We set the τ of 0.07, the learning rate of 1e-5, the training batch size of 800, and 10 epochs of training. In the experiment of LLM, we use **GLM2** to extract keyphrases. We set the text length to 1024 tokens, the learning rate to 1e-4, and the batch size to 48.

Baselines. To understand the effectiveness of **KESR**, we introduce the following relevance assessment methods for comparison. **BM25** [9]: An unsupervised method based on a probabilistic retrieval model. **Moco_relevance(Base)** [15]: is a Representation-based model trained using Moco architecture and contrastive learning. It is our base model and serves the main traffic of the search

Table 2: A comparison of performance for different models on rele-test dataset.

model	AUC	time(ms/ex)	mem(KiB/ex)
BM25	0.694	-	-
Moco_relevance(Base)	0.789	277	1.250
I3-Retriever	0.794	384	1.250
SBERT	0.759	384	1.250
SBERT w/ LLM-KPE	0.761	398	1.874
KESR	**0.840**	287	1.874
w/o RLDF	0.836	287	1.874
w/o LLM-KPE	0.821	287	1.874
w/ keyphrase concatenation	0.790	287	1.874
w/ keyphrase token pooling	0.786	287	1.874

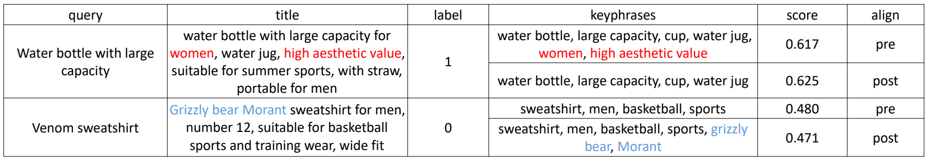
Table 3: The performance of BERTTagKPE and LLMs in key-test dataset.

Model	Presicion(%)	Recall(%)	F1(%)
BERTTagKPE	81.1	77.1	79.1
llama2-7b	79.8	78.9	79.4
intern2-7b	80.8	80.8	80.8
GLM2-6b (Unorder)	80.1	80.5	80.3
GLM2-6b	**80.5**	**81.2**	**80.9**

advertising system. **I3-Retriever** [3]: adopts a triple-encoder architecture and generates pseudo queries to reflect the query-title relationship. **SBERT** [8]: utilizes a shared BERT model, which is then fine-tuned with e-commerce data, to encode queries and titles separately. **SBERT w/ LLM-KPE**: utilizes the fine-tuned LLM(GLM2) to extract keyphrases, which are concatenated to ad titles in SBERT. **KESR**: To evaluate the contributions of the LLM-KPE module, RLDF method, and KERM module in KESR, we analyze the following variants. **KESR w/o LLM-KPE**: replaces the LLM-based keyphrase extraction with **BERTTagKPE**, which combines BERT+CRF with BIO sequence labeling. **KESR w/o RLDF**: removes the RLDF method to assess its impact. **KESR w keyphrase concatenation**: combines ad titles and keyphrases for title encoding, representing an alternative approach to keyphrase utilization. **KESR w keyphrase token pooling**: pools (MEAN) the keyphrase tokens from the ad title embeddings, representing another approach to keyphrase utilization.

4.2 MAIN RESULTS

The Superiority of Keyphrase-Enhanced Semantic Relevance Method. As shown in Table 2, KESR exhibits the best performance among all the methods in the relevance assessment task. Specifically, KESR surpasses Representation-based models such as BM25 by 21%, Base by 6.5%, I3-Retriever by 5.8%,



query	title	label	keyphrases	score	align
Water bottle with large capacity	water bottle with large capacity for women, water jug, high aesthetic value, suitable for summer sports, with straw, portable for men	1	water bottle, large capacity, cup, water jug, women, high aesthetic value	0.617	pre
			water bottle, large capacity, cup, water jug	0.625	post
Venom sweatshirt	Grizzly bear Morant sweatshirt for men, number 12, suitable for basketball sports and training wear, wide fit	0	sweatshirt, men, basketball, sports	0.480	pre
			sweatshirt, men, basketball, sports, grizzly bear, Morant	0.471	post

Fig. 3: Case studies on RLDF's impact on keyphrases. The red represents unique keyphrases before preference alignment, while the blue represents unique keyphrases after preference alignment.

and SBERT by 10.7% on AUC. These results demonstrate the effect of utilizing keyphrases for relevance modeling, refining and enhancing the semantics of ad titles, thereby achieving significant performance improvements. SBERT w/ LLM-KPE surpasses SBERT by 0.6% on AUC, which demonstrates the necessity of keyphrase extraction. In addition, KESR achieves an increase of 4.2%/6.3% compared to the KESR with keyphrase token pooling and the KESR with keyphrase concatenation. This result indicates the effectiveness of the way of utilizing keyphrases in KESR.

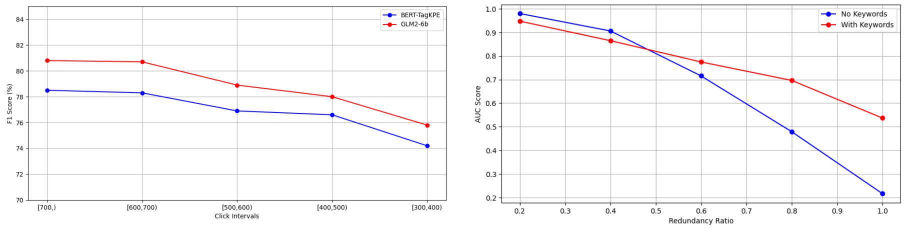
The Effect of LLM-based Keyphrase Extraction Method. Specifically, KESR improves AUC by 2.4% compared to KESR w/o LLM-KPE. This demonstrates that the extensive pre-trained knowledge of the LLM contributes to notable improvements in relevance assessment. Further, long-tail and case analyses are presented in the Effectiveness Analysis.

The Effect of RLDF Method. Furthermore, our proposed RLDF method demonstrates significant improvements in aligning upstream and downstream tasks. Specifically, KESR outperforms KESR w/o RLDF by 0.5% on AUC, Indicating its effectiveness in alleviating the discrepancy between the keyphrases generated by the LLM and those required by downstream relevance models.

Time and Memory Complexity. Considering industrial deployment, we evaluate the time and memory complexity under identical conditions. As shown in Table 2, incorporating keyphrases in KESR, results in a slight increase in inference time and storage requirements compared to the base model. However, this acceptable overhead doesn't affect the real-time online deployment and can bring about significant improvements in the model performance.

4.3 Effectiveness Analysis

Case Study. In Figure 3, for the sample labeled as 1 (relevant), the preference-aligned LLM removes gender-related terms (e.g., women, high aesthetic value), leading to an increase in the similarity score. Conversely, for the sample labeled as 0 (irrelevant), the preference-aligned LLM adds terms related to styles or categories (e.g., grizzly bear, Morant), resulting in a decrease in the similarity score. These results indicate that the RLDF method enables the LLM to accommodate user click feedback during keyphrase extraction. Consequently, the keyphrase-enhanced relevance model calculates more accurate relevance scores.



(a)The performance of different click intervals on keyphrase exaction

(b)the impact of information redundancy on relevance assessment

Fig. 4: An illustration of the impact of sample quality.

Quantitative Comparison. Table 3 compares the performance of different LLMs on the key-test dataset. It shows that all LLMs outperform BERTTagKPE in terms of F1, particularly when importance information is incorporated. This reflects the stronger language understanding and generation ability of the LLM.

Performance Comparison Across Click Intervals. In Figure 4a, we evaluate the performance of BERTTagKPE and GLM2-6b on ad keyphrase datasets with different click intervals. The results indicate that LLMs consistently outperform BERTTagKPE across all intervals, highlighting the superiority of LLMs in handling long-tail data.

Impact Analysis. In Figure 4b, we analyze the impact of keyphrases on AUC with varying levels of redundant information in ad titles. For positive examples, we define the redundancy rate as the proportion ρ of the words in the title that are not present in the query, relative to the length of the title. For negative examples, we compute the $1 - \rho$. We observed that, in the No Keyphrase setting, the Base model exhibits a sharp decline in performance as the redundancy rate increases. In contrast, in the With Keyphrase setting, the KESR model mitigates this downward trend. These findings suggest that incorporating keyphrases significantly improves relevance assessment performance, particularly in ad titles with higher redundancy.

Online Evaluation. We compare the model performance between KESR and previously deployed Base model through online A/B testing. Both models of the experiment are exposed to 5% of Taobao's search advertising traffic and run continuously for two weeks. KESR improves CTR for 0.96%, and CVR for 0.62%. This improvement can be attributed to the enhancement in relevance assessment ability, which pushes users to more desired products.

5 CONCLUSION

In this work, we propose a KESR framework for e-commerce relevance learning. KESR primarily consists of two components: a preference-aligned LLM for generating keyphrases needed by downstream relevance tasks, and a keyphrase-enhanced relevance model for assessing the relevance between queries and ad titles containing keyphrases. Experimental results demonstrate that KESR effectively enhances the performance of relevance assessment and user experience.

6 ACKNOWLEDGMENT

This research was supported by the National Key Research and Development Program of China (No.2024YFC3308300), the National Natural Science Foundation of China (No.U2133218), the Alibaba Group's Research Institute Internship Program and the Fundamental Research Funds for the Central Universities of China (No.FRF-MP-19-007 and No. FRF-TP-20-065A1Z).

References

1. Boudin, F., Gallina, Y., Aizawa, A.: Keyphrase generation for scientific document retrieval. arXiv preprint arXiv:2106.14726 (2021)
2. Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika **39**(3/4), 324–345 (1952)
3. Dong, Q., Liu, Y., Ai, Q., Li, H., Wang, S., Liu, Y., Yin, D., Ma, S.: Iˆ 3 retriever: Incorporating implicit interaction in pre-trained language models for passage retrieval. arXiv preprint arXiv:2306.02371 (2023)
4. Goldfarb, A., Tucker, C.: Online display advertising: Targeting and obtrusiveness. Marketing Science **30**(3), 389–404 (2011)
5. Huang, P.S., He, X., Gao, J., Deng, L., Acero, A., Heck, L.: Learning deep structured semantic models for web search using clickthrough data. In: Proceedings of the 22nd ACM international conference on Information & Knowledge Management. pp. 2333–2338 (2013)
6. Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.t.: Dense passage retrieval for open-domain question answering. arXiv preprint arXiv:2004.04906 (2020)
7. Nogueira, R., Cho, K.: Passage re-ranking with bert. arXiv preprint arXiv:1901.04085 (2019)
8. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019)
9. Robertson, S., Zaragoza, H., et al.: The probabilistic relevance framework: Bm25 and beyond. Foundations and Trends® in Information Retrieval **3**(4), 333–389 (2009)
10. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
11. Song, F., Yu, B., Li, M., Yu, H., Huang, F., Li, Y., Wang, H.: Preference ranking optimization for human alignment. arXiv preprint arXiv:2306.17492 (2023)
12. Song, M., Jing, L., Xiao, L.: Importance estimation from multiple perspectives for keyphrase extraction. arXiv preprint arXiv:2110.09749 (2021)
13. Tang, J., Wang, K.: Ranking distillation: Learning compact ranking models with high performance for recommender system. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 2289–2298 (2018)
14. Wang, Z., Zhao, L., Jiang, B., Zhou, G., Zhu, X., Gai, K.: Cold: Towards the next generation of pre-ranking system. arXiv preprint arXiv:2007.16122 (2020)
15. Zeng, Z., Huang, Y., Wu, T., Deng, H., Xu, J., Zheng, B.: Graph-based weakly supervised framework for semantic relevance learning in e-commerce. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. pp. 3634–3643 (2022)