CPRM: A LLM-based Continual Pre-training Framework for Relevance Modeling in Commercial Search

Kaixin Wu[1], Yixin Ji[1], Zeyuan Chen[1], Qiang Wang[2], Cunxiang Wang[3]
Hong Liu[1], Baijun Ji[4], Jia Xu[1]*, Zhongyi Liu[1], Jinjie Gu[1], Yuan Zhou[1], Linjian Mo[1]
[1]Ant Group, [2]Hithink RoyalFlush AI Research Institute
[3]Westlake University, [4]Soochow University
daniel.wkx@antgroup.com

Abstract

Relevance modeling between queries and items stands as a pivotal component in commercial search engines, directly affecting the user experience. Given the remarkable achievements of large language models (LLMs) in various natural language processing (NLP) tasks, LLM-based relevance modeling is gradually being adopted within industrial search systems. Nevertheless, foundational LLMs lack domain-specific knowledge and do not fully exploit the potential of in-context learning. Furthermore, structured item text remains underutilized, and there is a shortage in the supply of corresponding queries and background knowledge. We thereby propose **CPRM** (**C**ontinual **P**re-training for **R**elevance **M**odeling), a framework designed for the continual pre-training of LLMs to address these issues. Our CPRM framework includes three modules: 1) employing both queries and multi-field item to jointly pre-train for enhancing domain knowledge, 2) applying in-context pre-training, a novel approach where LLMs are pre-trained on a sequence of related queries or items, and 3) conducting reading comprehension on items to produce associated domain knowledge and background information (e.g., generating summaries and corresponding queries) to further strengthen LLMs. Results on offline experiments and online A/B testing demonstrate that our model achieves convincing performance compared to strong baselines.

1 Introduction

Relevance modeling is designed to evaluate the correlation between queries and items, an essential component of commercial search engines and crucial for the user experience. Mini-app service search is a common search application scenario. Unlike traditional e-commerce searches that only provide product search functions, mini-app services encompasses numerous scenarios such as livelihoods, government affairs, transportation, healthcare and dining. Moreover, the item consists of structured data with multiple fields; for instance, a hospital mini-app might include fields like title, keywords, category and description. Considering the diverse scenes and the complexity of structured data across multiple fields, conducting relevance modeling within the such search scenario poses a significant challenge. The current relevance model in commercial search systems is a semantic matching model, leveraging LLMs combined with domain-annotated data through supervised fine-tuning (SFT) methods. These LLMs like GPT-3 (Brown et al., 2020), GLM (Du et al., 2022), LLaMA (Touvron et al., 2023a), Qwen (Bai et al., 2023; Yang et al., 2024), having more extensive parameters and utilizing a massive corpora of texts during training compared to previous pre-trained models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019b) and XLNet (Yang et al., 2020), demonstrate superior performance in semantic matching tasks.

Despite the great success of LLMs, there still remain certain limitations in their application to relevance modeling. Firstly, LLMs are pre-trained on a broad range of data sources (Brown et al., 2020; Du et al., 2022; Touvron et al., 2023a,b), which do not afford special attention to particular domains (Wu et al., 2023; Cui et al., 2023; Xiong et al., 2023), resulting in a lack of domain-specific knowledge. Besides, queries tend to be colloquial and present in short-text form, whereas items are typically expressed in a more formal long-text form, leading to a "semantic gap" (Lian et al., 2019; Qi et al., 2020; Kumar and Sarkar, 2021) between their representations. Secondly, the pre-training phase of LLMs is "task-agnostic" (Brown et al., 2020), which impedes direct connection with downstream tasks and precludes the possibility of in-context pre-training enhancements tailored for these tasks (Min et al., 2022; Gu et al., 2023). Finally, the item tends

* Corresponding author.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Industry Track), pages 998–1008

to be highly structured and difficult to leverage, which prevents LLMs from fully realizing their potential with such data.

To address the above problems, we investigate a **C**ontinual **P**re-training approach of LLMs for **R**elevance **M**odeling, CPRM for short. Initially, we introduce a pre-training technique using pairs of queries and multi-field item as inputs. This method enables the LLMs to explicitly model the semantic representations between queries and items, thus bridging the semantic gap between them. Subsequently, we collect sets of semantically similar queries and items based on user click logs, then further refine these sets through semantic modeling to filter out semantically irrelevant cases. Following that, we reorder these refined sets according to semantic similarity and ultimately construct in-context pre-training instances via prompting techniques. Utilizing this approach to data construction, LLMs are able to make better predictions within such contexts during the training process, which benefits the efficient learning current domain knowledge for LLMs. Lastly, we employ a larger parameter LLM (teacher LLM) to conduct reading comprehension on structured item data, facilitating the generation of relevant domain knowledge for pre-training, which can be considered as a secondary development and exploitation of the item. More specifically, we leverage teacher LLM to summarize and paraphrase item to produce fluent domain knowledge, while also guiding teacher LLM to produce background information related to the items. Additionally, we prompt teacher LLM to create diverse queries and provide further reasons for their generation. In summary, the contributions of this paper are as follows:

- To our knowledge, we are the first to systematically propose a continual pre-training approach of LLMs specifically designed for relevance modeling tasks.

- We propose a CPRM framework with three components. Firstly, the joint pre-training of queries and multi-field item to enhance domain knowledge of LLMs. Secondly, in-context pre-training by constructing collections of semantically similar queries or items. And thirdly, reading comprehension of structured items employed to strengthen the capabilities of LLMs further.

- Our approach has been validated on real-world industry data, outperforming strong baselines significantly in both offline experiments and online A/B testing.

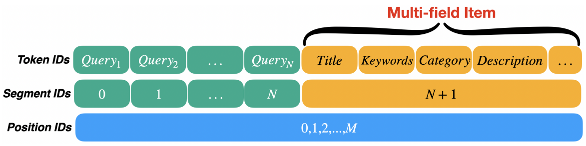

Figure 1: Joint queries and multi-field item for pre-training. An example of the mini-app search scenario.

2 Related Work

Relevance modeling in search corresponds to the semantic matching task in NLP. With the advancement of neural network and pre-trained models, deep semantic matching models have become mainstream. Deep semantic matching models are categorized into two types: representation-based (Shen et al., 2014; Palangi et al., 2015; Rao et al., 2019) and interaction-based methods (Chen et al., 2016; Hu et al., 2014; Pang et al., 2016; Parikh et al., 2016). The former focuses on learning low-dimensional representations, while the latter emphasizes capturing the interactions between inputs. The representation-based model with independently encoded inputs struggles to capture complex correlations, whereas interaction-based methods that concatenate the two inputs for semantic computation can alleviate this issue.

In recent years, pre-trained models like BERT (Devlin et al., 2019) has show its superiority on natural language understanding (NLU) tasks. Consequently, both representation-based and interaction-based methods have begun leveraging the capabilities of these pre-trained models for semantic modeling. Most recently, LLMs like GPT-3 (Brown et al., 2020), GLM (Du et al., 2022), LLaMA (Touvron et al., 2023a), Qwen (Bai et al., 2023; Yang et al., 2024) pre-trained extensive volumes of data with numerous parameters have garnered significant performance in language understanding, generation and reasoning tasks. Compared to traditional pre-trained models like BERT (Devlin et al., 2019), LLMs possess significant advantages in both the scale of pre-training data and the quantity of model parameters, leading to their evident superiority in performance across a variety of downstream tasks. Recent research work (Sun et al., 2023; Spatharioti et al., 2023; Zhu et al., 2024) indicates that combining LLMs with downstream applications presents significant potential, LLMs can achieve competitive or even superior results compared to traditional supervised methods on information retrieval benchmarks. Some research leverage LLMs for relevance modeling in search engines, adopting approaches such as behavior-augmented (Chen et al., 2023, 2024) or robust learning (Liu et al., 2024) to improve the capability of relevance modeling. Our work mainly focuses on enhancing LLMs from the perspective of continual pre-training for relevance modeling. LLMs are pre-trained on a wide variety of data sources (Brown et al., 2020; Du et al., 2022; Touvron et al., 2023a,b) without pay more attention on specific domains, resulting in a lack of domain knowledge. On the other hand, the pre-training phase of LLMs is task-agnostic (Brown et al., 2020), making it difficult to direct connect with downstream tasks. This also means we can't easily customize the pre-training process to better fit those tasks (Min et al., 2022; Gu et al., 2023). Previous work of injecting domain knowledge involves continued training of pre-trained models on domain-specific data (Gururangan et al., 2020; Shi et al., 2023), as well as incorporating knowledge graphs (Liu et al., 2019a) or selectively masking important information (Gu et al., 2020; Xu et al., 2023; Sanyal et al., 2023; Zhou et al., 2023). Another line of research aims to enhance the pre-training for downstream tasks, which simple concatenate relevant documents together for in-context pre-training (Min et al., 2022; Gu et al., 2023; Shi et al., 2024). However, these approaches assume that downstream tasks contain only a single domain representation, neglecting the possibility of there being multiple or more. For instance, in commercial search relevance tasks, queries and items belong to two distinct domains with substantial differences. Our research work involves injecting domain knowledge and conducting in-context pre-training simultaneously, while being able to establish a connection between the two domains.

3 Problem Formulation

Given a query Q and an item I, LLM-based relevance modeling in search engines aims to predict the relevance degree between them. Essentially, referring to PET (Schick and Schütze, 2021), we first design the prompt $\mathrm{P}(Q, I)$, and LLM determines which verbalizer v (e.g., "no" or "yes") is the most likely candidate for "[Mask]" conditioned on the likelihood $\mathrm{M}(v|\mathrm{P}(Q, I))$. The above process is defined as follows:

$$\mathrm{P}(Q, I) = Is\ [Q]\ and\ [I]\ related?\ \text{[Mask]} \quad (1)$$

$$y = \mathrm{M}(v \mid \mathrm{P}(Q, I)), \quad \text{for } v \in \{\text{no, yes}\}, \quad (2)$$

where relevance label $y \in \{0, 1\}$ can be associated with a verbalizer (e.g., "no" or "yes") from the vocabulary of LLMs to represent "irrelevant" and "relevant" between Q and I respectively. To enable the adaptation of general LLMs to the relevance modeling task, SFT operation is selected for training with the cross-entropy loss function. Consequently, the relevance degree could be given from LLMs for subsequent applications in search scenarios.

4 Methodology

In this section, we present the details of the CPRM framework, including Domain Knowledge Enhancement (DKE), In-Context Pre-training (ICP) and Reading Comprehension Distillation (RCD).

4.1 Domain Knowledge Enhancement (DKE)

Different from conventional pre-training methods, we jointly pre-train the structured item data with multiple queries as shown in Figure 1. Each item encompasses multiple fields, including title, keywords, category, description, etc., with the query being the most frequently searched top query for a given item. For each query or item, we employ segment embeddings to distinguish between different texts. For convenience, we add special tokens "<|startofpiece|>" and "<|endofpiece|>" between the queries and item as segment embeddings to differentiate them. Furthermore, queries and item are combined for position encoding, thereby allowing LLMs to explicitly model the relationships between them during pre-training process. Due to constraints on response time for online services, when calculating relevance scores between queries and items, only a limited number of item fields (such as title and keywords) are considered. Consequently, domain knowledge from other unused item fields, such as description, can be incorporated through continual pre-training. Considering that relevance modeling is a NLU task, we adopt both token-level masked language modeling (MLM) (Devlin et al., 2018) and segment-level MLM pre-training strategies for LLMs. Therefore, the optimization objective is:

$$\mathcal{L}(\theta) = \min_{\theta} \alpha \mathcal{L}_{\text{t-MLM}}(\theta) + (1-\alpha) \mathcal{L}_{\text{s-MLM}}(\theta), \quad (3)$$

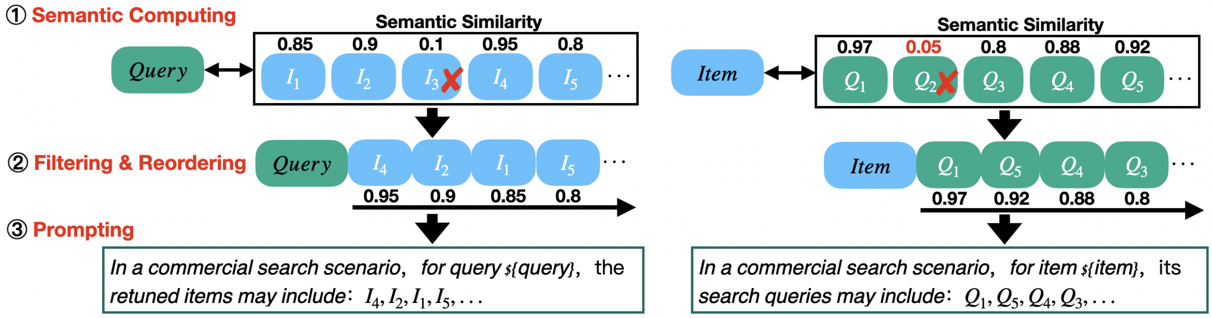

Figure 2: In-context pre-training instances construction. The left and right figures represent the ICP instances constructed from similar item sets and similar query sets respectively.

where θ is the parameters of the model, $\mathcal{L}_{\text{t-MLM}}(\theta)$ and $\mathcal{L}_{\text{s-MLM}}(\theta)$ represent token-level MLM loss and segment-level MLM loss respectively, we set $\alpha = 0.7$ in our experiments.

4.2 In-Context Pre-training (ICP)

We construct in-context pre-training instances using historical click logs from real-world business search scenario. The overall idea is to build collections of semantically similar queries and items as pre-training data to further stimulate in-context learning capabilities of LLMs. The detailed data construction methodology is as follows:

Coarse Screening. Utilizing the click logs, we establish a mapping from Queries to Items (denoted as $Q2I$ and from Items to Queries (denoted as $I2Q$), sorting them by the number of clicks in descending order. Consequently, within the $Q2I$ mapping, for a query $Query$ there is an associated set of items $I^{Query} = \{I_1, I_2, ..., I_N\}$. These items can be considered as a preliminarily semantically related collection under the specific constraint of query $Query$. Vice versa for $I2Q$ mapping.

Fine Screening. Following described above, cases may be introduced that received clicks but are semantically unrelated. We employ *Contriever* (Izacard et al., 2022), a semantic model, to encode text into vectors, and then calculate the similarity between various text representations for semantic filtering. For set I^{Query}, when the following condition is met:

$$Sim(Query, I_k) < \sigma, \quad \text{for } k \in [1, N], \quad (4)$$

it signifies that $Query$ and I_k are semantically unrelated and require filtering, where $Sim(\cdot)$ is similarity function, σ is a threshold. **Data Construction.** As shown in the left of Figure 2, we subsequently obtain a collection of items that are semantically relevant to the query, then sort these items by semantic similarity in ascending order. Finally, we concatenate the query with the sorted collection of items to create an ICP instance via prompting. The right of Figure 2 illustrates how to construct ICP instances under $I2Q$ mapping, namely, obtaining a set of semantically related queries given the constraint of an item.

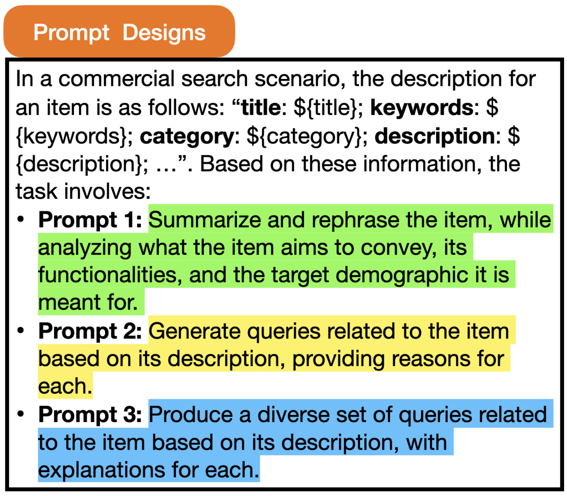

Figure 3: Prompt for reading comprehension on item.

Why adopt this construction manner? By assembling collections of items under a specified query or collections of queries under a specified item, LLMs can make better predictions based on the context during the pre-training process, enabling more efficient learning within the current domain. Moreover, the reordering in ICP instances implicitly indicates the strength of relevance between queries and items, enabling LLMs to model the degree of their association effectively. On the other hand, by linking queries and items in our ICP instances, we enable LLMs to model their semantic representations directly.

4.3 Reading Comprehension Distillation (RCD)

We employ the teacher LLM for reading comprehension on items, with the prompt design shown in Figure 3. Assume that in a mini-app search scenario, we need to provide the mini-app's structured information like title, keywords, category, and description, and utilizing prompt template instructions to invoke teacher LLM. This generates the reading comprehension pre-training instances.

Why design the prompt in this way? We have several reasons for this design choice. Firstly, item text is structured data and difficult to utilize, lacking in relevant background knowledge. Through summarizing and rephrasing with Prompt 1, fluent domain knowledge can be generated. Additionally, understanding and analyzing items can instruct teacher LLM in generating relevant background knowledge. Secondly, by using Prompt 2 and Prompt 3 guide teacher LLM to generate related and diversified queries, enriching the supply of suitable queries. We also instruct teacher LLM to provide further explanations for the generated queries. This approach not only facilitates the generation of relevant domain knowledge but also allows downstream models to significantly improve their understanding and handling of the item when utilizing these data. Employing teacher LLM for reading comprehension on items can be considered a secondary development and utilization of item data, enriching the domain knowledge further. Pre-training LLMs on the above data can also be seen as a process of knowledge transfer from teahacher LLM to LLMs.

5 Experiments

5.1 Experimental Settings

Dataset & Evaluation Metrics. We utilize the real-world mini-app search scenario data for verification. The pre-training data includes three parts: DKE data (4M), ICP data (4M) and RCD data (500K). The first part is sampled from the mini-app search scenarios and consist of structured items containing multiple fields. For top 500K most frequently visited items, we sample 5 top queries based on click logs for each item, which are then concatenated with multi-field item to serve as pre-training examples for adapting relevance tasks. The second part is in-context pre-training data, where we construct these examples based on the real-world search click logs using the method described in Section 4.2, and subsequently randomly sample 4M from them. The third part is reading comprehension data, for which we utilize teacher LLM (e.g. Qwen2-72B (Yang et al., 2024)) to perform reading comprehension on item data. The SFT data consists of three parts: train set (625K), valid set (35K) and test set (35K). These data are sourced from real mini-app search results and then are generated through manual annotation. The human-annotated data for relevance tasks are in format of triples <*Query*, *Item*, *Label*>, the data statistics as shown in Table 1. The annotated data have only two levels of relevance: "*#Relevant*" and "*#Irrelevant*". For evaluation, we employ three widely used metrics Acc., F1 and AUC to evaluate model performance, with higher values indicating better performance. Note that AUC serves as the most important metric in relevance tasks while the others provide auxiliary supports for our analysis.

Dataset	#Sample	#Query	#Item	#Relevant	#Irrelevant
Train	625,292	92,711	32,219	370,887	254,405
Valid	35,252	5,016	8,250	20,023	15,229
Test	35,057	5,426	8,406	19,164	15,893

Table 1: Data statistics (# of numbers)

Baselines. We selected classic NLU-based models and commonly used LLMs as our baseline models: DSSM (Shen et al., 2014), ReprBERT (Yao et al., 2022), BERT (Devlin et al., 2018), GLM (Du et al., 2022), Qwen2 (Yang et al., 2024), ChatGPT & GPT-4 (Team, 2024).

Implementation Details. All our pre-training experiments are conducted on the GLM-2B. The model configuration set to 36 layers, hidden size of 2048, FFN size of 8192 and 32 attention heads. We utilize adam (Kingma and Ba, 2017) optimizer and the warmup steps and learning rate set 28K and e^{-4}. All models are pre-trained on 8 A100 GPUs for 2 epochs and the batch size set 64. During SFT, all models are trained for 5 epochs on 8 A100 GPUs and the batch size is 8. The adam optimizer is employed and warmup steps and learning rate set to 5K and $2e^{-5}$ respectively. When constructing the ICP instances, we utilize facebook's open-source multilingual *Contriever*[1] (Izacard et al., 2022) model for semantic filtering.

5.2 Offline Experimental Results

Performance Comparison. Table 2 presents the performance of different baselines and various con-

[1]https://github.com/facebookresearch/contriever

#	Model	**Acc.** (%)	Δ_{Acc}	**F1** (%)	Δ_{F1}	**AUC** (%)	Δ_{AUC}
	Only fine-tuning on supervised datasets						
1	DSSM	70.32	-	71.21	-	70.64	-
2	ReprBERT	80.65	-	82.77	-	80.24	-
3	BERT-Base (0.1B)	82.24	-	84.33	-	81.79	-
4	BERT-Large (0.3B)	83.47	-	85.65	-	83.01	-
5	Qwen2-0.5B	80.48	-	79.52	-	81.63	-
6	Qwen2-1.5B	91.17	-	92.08	-	90.90	-
7	GLM-0.3B	85.93	-	87.32	-	85.64	-
8	GLM-2B	91.16	-	91.95	-	91.04	-
9	GLM-5B	93.53	-	94.12	-	93.41	-
10	GLM-10B	**93.70**	-	<u>94.26</u>	-	**93.62**	-
	Without fine-tuning						
11	ChatGPT (+8-shot)	62.93	-	59.78	-	64.88	-
12	GPT-4 (+8-shot)	61.89	-	67.44	-	60.83	-
	Continual pre-training LLM and then fine-tuning on supervised datasets						
13	GLM-2B	91.16	+0.00	91.95	+0.00	91.04	+0.00
14	+ DKE	92.28	+1.12	92.99	+1.04	92.15	+1.11
15	+ ICP	92.72	+1.56	93.40	+1.45	92.58	+1.54
16	+ RCD	91.59	+0.43	92.36	+0.41	91.46	+0.42
17	+ DKE + ICP	93.33	+2.17	93.93	+1.98	93.23	+2.19
18	+ DKE + ICP + RCD (a.k.a. **CPRM**)	<u>93.64</u>	+2.48	**94.42**	+2.47	<u>93.49</u>	+2.45

Table 2: Performance of different baselines and various continual pre-training models on the relevance task. **Bold** and <u>underline</u> represent the best and second best result respectively. Improvements over variants are statistically significant with $p < 0.05$.

tinual pre-training models on the relevance task. From the experimental results, GLM demonstrates strong competitiveness, achieving superior performance even with similar parameter numbers compared to BERT-Large (line 7 vs. line 4). We also conducts experiments on the GLM of various parameter sizes, and the results show that as the model size increases, its performance gradually improves. However, with further increases in model size, the performance gains become progressively smaller. Specifically, GLM-10B achieves only a 0.21% improvement in AUC over GLM-5B (line 10 vs. line 9). Compared to other latest LLMs such as Qwen2, our GLM also shows impressive performance at a similar parameter scale (line 7 vs. line 5, line 8 vs. line 6). ChatGPT & GPT-4 performed poorly compared to other SFT-based baseline systems; this is because the task data belongs to a proprietary domain, and models without SFT operations have relatively poor discriminative ability. We conducts continued pre-training experiments on GLM-2B, and the experimental results demonstrate that all three different methods result in performance improvements compared to the baseline model. Notably, the DKE and ICP methods achieve significant performance enhancements, with respective gains of 1.11% and 1.54% in AUC. This is because both methods are constructed based on domain-specific data and jointly training semantically related queries or items can further enhance model performance. The experimental results also indicate that integrating different continued pre-training methods can further strength model performance (line 17 and line 18), with the combination of all three methods leading to the greatest performance gain, making it comparable to that of GLM-10B (line 18 vs. line 10). Our CPRM model achieves the highest F1 score (94.42%) among all models compared.

Analysis on Query Length. We compare the performance of different models at various query lengths on test set. As shown in Figure 4, our CPRM model outperforms the baseline model across all length intervals, especially on longer queries (when the length greater than 15), where the CPRM model demonstrates a significant improvement performance gains, surpassing the baseline model by 15.85% in AUC (92.27% vs.

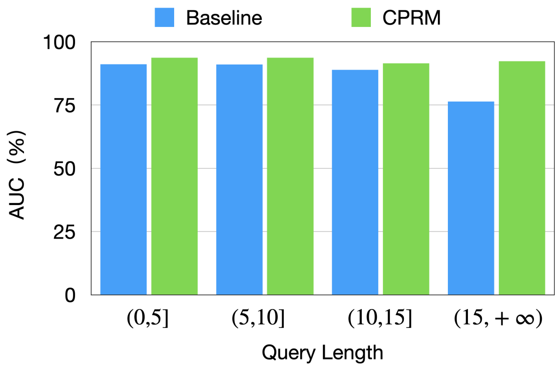

Figure 4: Performance of different query lengths.

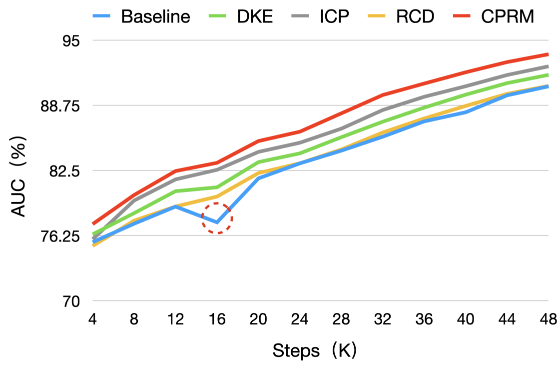

Figure 5: Performance of different training steps.

76.42%). This suggests that the CPRM model possesses a superior ability to understand and deal with long queries. We speculate that this advantage may be attributable to the ICP and RCD methods. Since the ICP method semantically aggregates historical search queries, allowing the model have the possibility encountered related long queries and to understand their semantics in the in-context pre-training process. On the other hand, the RCD method generates diverse queries, thereby enriching the model's understanding of various long query types.

Impact of Training Steps. As shown in Figure 5, we compare models' performance with different pre-training methods across various training steps. The experimental results show that models trained with all three different pre-training methods surpasse the baseline across various training steps. The CPRM model, which combines all three methods, achieves the best performance at each step. These evidence highlights the robustness of our proposed approach. Interestingly, an phenomenon observed from the figure is that the baseline model's performance significantly decreases at the 16K training step before it gradually increases thereafter. The reason is due to the significant difference between the current task data and the data previously seen by the LLM, resulting in challenges for the LLM in fitting this domain-specific data. None of the other pre-training methods exhibits this phenomenon; instead, the performance of these models steadily improved at each training step. This indicates that our proposed methods are beneficial for the domain adaptation of LLM.

6 Online A/B Testing

We deploy the proposed model on the online search platform to provide search services for mini-apps, and conduct a two-week online A/B testing with 5% proportion of the experiment traffic. The experimental results show that, compared to the baseline system (GLM-2B), our CPRM method yields a statistically significant increase of 0.32% in valid PVCTR[2] at a 95% confidence level. Human evaluation indicates a 0.75% reduction in Badcase@10 metric and a 4.71% decrease in the Error Filtering Rate[3]. The model has now been serving search functions to mini-apps for over nine months. These results suggest that our proposed method can effectively enhance relevance models' performance in real-world search systems.

7 Conclusion

In this paper, we have investigated CPRM framework, a continued pre-training approach of LLMs tailored to relevance modeling tasks, which comprises three methods: DKE, ICP and RCD. Both offline experiments and online A/B testing results demonstrate that our proposed method boosts the search relevance of LLMs effectively. Our model has been successfully deployed online search platform.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang

[2]Page view click-through rate, the number of valid clicks divided by the number of searches.

[3]The number of relevant items that are incorrectly filtered out divided by the total number of filtered items.

Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2016. Enhanced lstm for natural language inference. *arXiv preprint arXiv:1609.06038*.

Zeyuan Chen, Wei Chen, Jia Xu, Zhongyi Liu, and Wei Zhang. 2023. Beyond semantics: Learning a behavior augmented relevance model with self-supervised learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4516–4522.

Zeyuan Chen, Haiyan Wu, Kaixin Wu, Wei Chen, Mingjie Zhong, Jia Xu, Zhongyi Liu, and Wei Zhang. 2024. Towards boosting llms-driven relevance modeling with progressive retrieved behavior-augmented prompting.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. Pre-training to learn in context.

Yuxian Gu, Zhengyan Zhang, Xiaozhi Wang, Zhiyuan Liu, and Maosong Sun. 2020. Train no evil: Selective masking for task-guided pre-training.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks.

Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. *Advances in neural information processing systems*, 27.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Lakshya Kumar and Sagnik Sarkar. 2021. Neural search: Learning query and product representations in fashion e-commerce.

Yijiang Lian, Zhijie Chen, Jinlong Hu, Kefeng Zhang, Chunwei Yan, Muchenxuan Tong, Wenying Han, Hanju Guan, Ying Li, Ying Cao, Yang Yu, Zhigang Li, Xiaochun Liu, and Yue Wang. 2019. An end-to-end generative retrieval method for sponsored search engine –decoding efficiently into a closed target domain.

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*, pages 4718–4725.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2019a. K-bert: Enabling language representation with knowledge graph.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. Metaicl: Learning to learn in context.

H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, X. Song, and R. Ward. 2015. Semantic modelling with long-short-term memory for information retrieval.

Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. 2016. Text matching as image recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training.

Jinfeng Rao, Linqing Liu, Yi Tay, Wei Yang, Peng Shi, and Jimmy Lin. 2019. Bridging the gap between relevance matching and semantic matching for short text similarity modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5370–5381.

Soumya Sanyal, Yichong Xu, Shuohang Wang, Ziyi Yang, Reid Pryzant, Wenhao Yu, Chenguang Zhu, and Xiang Ren. 2023. Apollo: A simple approach for adaptive pretraining of language models for logical reasoning.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze questions for few shot text classification and natural language inference.

Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014. Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 23rd international conference on world wide web*, pages 373–374.

Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Gergely Szilvasy, Rich James, Xi Victoria Lin, Noah A. Smith, Luke Zettlemoyer, Scott Yih, and Mike Lewis. 2024. In-context pretraining: Language modeling beyond document boundaries.

Zhengxiang Shi, Francesco Tonolini, Nikolaos Aletras, Emine Yilmaz, Gabriella Kazai, and Yunlong Jiao. 2023. Rethinking semi-supervised learning with language models.

Sofia Eleni Spatharioti, David M. Rothschild, Daniel G. Goldstein, and Jake M. Hofman. 2023. Comparing traditional and llm-based search for consumer choice: A randomized experiment.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agent. *arXiv preprint arXiv:2304.09542*.

OpenAI Team. 2024. Gpt-4 technical report.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task.

Yan Xu, Mahdi Namazifar, Devamanyu Hazarika, Aishwarya Padmakumar, Yang Liu, and Dilek Hakkani-Tür. 2023. Kilm: Knowledge injection into encoder-decoder language models.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. Xlnet: Generalized autoregressive pretraining for language understanding.

Shaowei Yao, Jiwei Tan, Xi Chen, Juhao Zhang, Xiaoyi Zeng, and Keping Yang. 2022. Reprbert: Distilling bert to an efficient representation-based relevance model for e-commerce. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4363–4371.

Wangchunshu Zhou, Ronan Le Bras, and Yejin Choi. 2023. Commonsense knowledge transfer for pre-trained language models.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zhicheng Dou, and Ji-Rong Wen. 2024. Large language models for information retrieval: A survey.

A Baselines

We compare our proposed CPRM model with the following baselines:

- **DSSM** (Shen et al., 2014) is a classic two-tower structure text matching model that constructs representations for the query and item independently, using cosine similarity to measure their relevance.
- **ReprBERT** (Yao et al., 2022) is a representation-based BERT model that utilizes novel interaction strategies to balance performance and latency.
- **BERT** (Devlin et al., 2018) has achieved great success on NLP tasks as an interaction-based model. Here, we concatenate the query and item as the model input for relevance modeling.
- **GLM** (Du et al., 2022) is a powerful LLM architecture with various parameter sizes to suit different business scenarios. Our LLM online system is developed based on the GLM, thus all our experiments are mainly conducted on the GLM.
- **Qwen2** (Yang et al., 2024) is currently one of the newest and the state-of-the-art (SOTA) open-source LLMs for Chinese NLP tasks.
- **ChatGPT**[4] **& GPT-4** (Team, 2024) are the SOTA closed-source LLMs. We employ the direct generation approach for relevance task evaluation.

B Model Deployment

LLMs have achieved significant performance improvements in relevance tasks, but their large parameter size leads to low inference efficiency, thus affecting their deployment online. We have designed a solution that allows for the real-time use of LLMs' relevance scores. As shown in Figure 6, the online relevance model for search consists of two parts: the GLM-0.3B model serves as the online model to respond to search queries in real-time, while the GLM-2B model employ a T+1 update strategy to score historical Q-I pairs and cache them offline. The online relevance service gives priority to using the cached scores from GLM-2B; if the cache does not exist, it calls on the GLM-0.3B online model. Currently, using GLM-2B's offline caching scoring has covered over 60% of mini-app search requests, significantly alleviating the request pressure on the online model.

[4]The version is GPT-3.5-turbo.

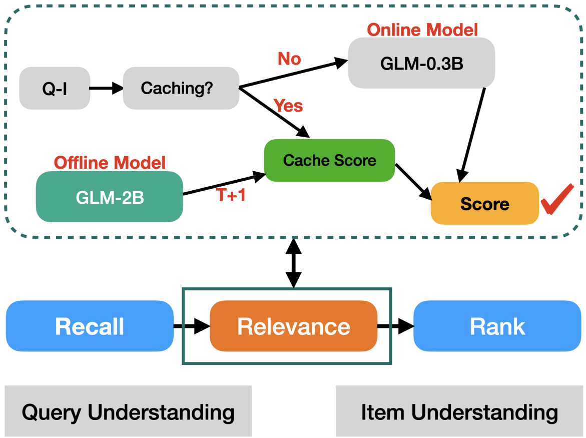

Figure 6: Deployment of the CPRM relevance model.

C Case Study

As shown in Figure 7, we provide several cases to compare the relevance output results between the baseline (GLM-2B) and CPRM models. From these cases, we can observe that the CPRM method is able to supplement additional domain knowledge to correct erroneous prediction results. Furthermore, CPRM demonstrates a stronger understanding of long and complex queries (query length greater than 15) compared to the baseline. The SFT data format can be referred to in Figure 8.

Case 1

- **Query**: 租无人机航拍 (Rent drone for aerial photography)
- **Item**: 标题:宜租机租手机 关键词:宜租机租手机手机租赁,租苹果14手机,免押金租手机,租全新手机,租手机 (Title: Yizuji rents mobile phones Keywords: Rent a phone from Yizuji, lease an iPhone 14, rent a phone without a deposit, rent a brand new phone, rent a phone)
- **Ground Truth**: 相关 (relevant)
- **GLM-2B**: 不相关 (irrelevant)✘
- **CPRM**: 相关 (relevant)✔

Pre-training Data in CPRM

- **Multi-field Item**: 标题:宜租机租手机 类目:二手/租赁,消费品租赁,其他租赁 关键词:宜租机租手机手机租赁,租苹果14手机,免押金租手机,租全新手机,租手机 描述:宜租机租手机是一种可以让用户在线租笔记本,租平板, 租手机, 租电脑, 租无人机, 租游戏机, 租相机, 租耳机, 租手表, 租游戏本等多种数码产品的平台, 在这里你可以不用花多少钱就能体验各种新机, 苹果、小米、华为等多种品牌, 应有尽有, 快来选择一款心仪的手机试试吧！ (Title: Yizuji rents mobile phones **Category**: Second-hand/rental, consumer goods rental, other rental Keywords: Rent a phone from Yizuji, lease an iPhone 14, rent a phone without a deposit, rent a brand new phone, rent a phone **Description**: **Yizuji is a platform that allows users to rent a variety of digital products online, including laptops, tablets, mobile phones, computers, drones, game consoles, cameras, headphones, watches, and gaming laptops**. Here, you can experience a variety of new devices from brands such as Apple, Xiaomi, Huawei, and more without spending a lot of money. Come and choose your favorite phone to try out!)

Case 2 - Long and complex query

- **Query**: 国泰CES半导体芯片行业ETF联接C (Guotai semiconductor chip industry ETF connection C)
- **Item**: 标题:国泰基金资产证明 关键词:资产证明,国泰基金国泰基金,资产证明 (Title: Guotai fund asset certification Keywords: Asset certification, Guotai fund)
- **Ground Truth**: 不相关 (irrelevant)
- **GLM-2B**: 相关 (relevant)✘
- **CPRM**: 不相关 (irrelevant)✔

Case 3 - Long and complex query

- **Query**: 广发道琼斯美国石油指数(QDII-LOF)C 004243 (Guangfa Dow Jones U.S. oil index (QDII-LOF) C 004243)
- **Item**: 标题:大成基金猜涨跌 关键词:上证综指,大成基金,猜指数,猜大盘,猜涨跌 (Title: Dacheng Fund's Speculation on Bullish Market Keywords: Shanghai composite Index, Dacheng fund, predicting index, predicting market, bullish or bearish predictions)
- **Ground Truth**: 相关 (relevant)
- **GLM-2B**: 不相关 (irrelevant)✘
- **CPRM**: 相关 (relevant)✔

Figure 7: Case study.

Example 1

- **Query**: 电动车缴费 (Electric vehicle payment)
- **Item**: 标题:交通123违章查询缴纳-车易通 关键词:查询违章,交通违法查询,违法查询,交通违章查询,违章代办 (Title: Traffic 123 violation inquiry and payment-Cheyitong Keywords: Violation inquiry, traffic violation inquiry, violation search, traffic violation inquiry, violation agency)
- **Label**: 相关 (relevant)

Example 2

- **Query**: 个人社保余额查询 (Personal social security balance inquiry)
- **Item**: 标题:随申办 关键词:医保,居保,户口,学区,出入境随申办 (Title: Suishenban Keywords: Medical insurance, residential insurance, household registration, school district, Suishenban for entry and exit)
- **Label**: 相关 (relevant)

Example 3

- **Query**: 春季消费领积分 (Earn Points from spring shopping)
- **Item**: 标题:领取积分兑换商品 关键词:积分兑换商品中心,商城积分兑换商品,积分兑换商品商城,商家积分兑换好礼,积分兑换中心 (Title: Earn points to redeem products Keywords: Points redemption product center, mall points redemption products, points redemption mall, merchant points gift redemption, points redemption center)
- **Label**: 不相关 (irrelevant)

Example 4

- **Query**: 手机充值联通 (Mobile phone recharge Unicom)
- **Item**: 标题:中国电信流量卡营业厅豪斯莱 关键词:流量卡,电话卡,移动手机卡,移动电话卡,抖音流量卡 (Title: China Telecom data card business hall Houslai Keywords: Data card, phone card, mobile phone card, mobile phone card, Douyin data card)
- **Label**: 不相关 (irrelevant)

Figure 8: SFT data examples.