# BSharedRAG: Backbone Shared Retrieval-Augmented Generation for the E-commerce Domain

**Kaisi Guan    Qian Cao    Yuchong Sun    Xiting Wang** *    **Ruihua Song** *

Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China

{guankaisi,caoqian4real,ycsun,xitingwang,rsong}@ruc.edu.com

## Abstract

Retrieval Augmented Generation (RAG) system is important in domains such as e-commerce, which has many long-tail entities and frequently updated information. Most existing works adopt separate modules for retrieval and generation, which may be suboptimal since the retrieval task and the generation task cannot benefit from each other to improve performance. In this paper, we construct a high-quality e-commerce dataset and use e-commerce as an example domain to show how to ensure effective transfer between the retrieval and generation tasks. We propose a novel Backbone Shared RAG framework (**BSharedRAG**). It first uses a domain-specific corpus to continually pre-train a base model as a domain-specific backbone model and then trains two plug-and-play Low-Rank Adaptation (LoRA) modules based on the shared backbone to minimize retrieval and generation losses respectively. Experimental results indicate that our proposed BSharedRAG outperforms baseline models by 5% and 13% in Hit@3 upon two datasets in retrieval evaluation and by 23% in terms of BLEU-3 in generation evaluation. Our codes, models, and dataset are available at https://bsharedrag.github.io.

## 1 Introduction

Large Language Models (LLMs) have shown impressive performance in reasoning and remembering extensive knowledge (Radford et al., 2019; Brown et al., 2020; Achiam et al., 2023; Anthropic, 2024; Zhao et al., 2023b; Touvron et al., 2023a,b; Yang et al., 2023; Chen et al., 2024). However, when facing domain-specific questions, especially the e-commerce domain that has many long-tail entities and frequently updated information, LLMs often have issues such as hallucinations (Shi et al., 2023c; Wang et al., 2023; Zhao et al., 2023a). Retrieval-Augmented Generation

(RAG) has been proposed as an effective method for such cases (Zhu et al., 2024; Gao et al., 2024; Zhao et al., 2024). However, previous works (Zhang et al., 2023; Shi et al., 2023b; Lin et al., 2023; Zhang et al., 2024) mainly adopt a general-domain retriever and generator directly, which may be suboptimal due to the lack of domain-specific knowledge. How to construct a domain-specific RAG system is an important research problem. To answer this question, we need to solve the following two challenges.

*C1 (dataset)*: How to build a high-quality RAG dataset with an informative knowledge base and correct question-answer (QA) pairs? For important domains like e-commerce, the existing knowledge bases are usually short and noisy, as most of them are from user-generated product reviews. The QA pairs are also of various qualities (Deng et al., 2023) and may contain subjective or conflicting information because the answers are provided by users.

*C2 (framework)*: How to design a suitable RAG framework for domain adaptation? As shown in Figure 1 (a), existing works typically adopt separate modules for retrieval and generation (Shi et al., 2023b; Lin et al., 2023; Zhang et al., 2023; langchain, 2023; Liu, 2022), for example, using BERT-like (Devlin et al., 2019) embedding models for retrieving documents and an LLM for generating answers. In this design, retrieval and generation operate independently, hindering their potential to achieve optimal performance. It is often challenging for the retriever to leverage the continuous pre-training of the generator, while the generator's performance may be constrained by retrieved documents that, although similar to the query, are less effective in addressing the questions (Section 5.2.3). Although we can build the retriever and generator both on a fully shared LLM and optimize it with multi-task learning as shown in Figure 1 (b), it can suffer from negative transfer between tasks, *i.e.*, performance decrease due to the potential conflicts

---

* Corresponding authors.

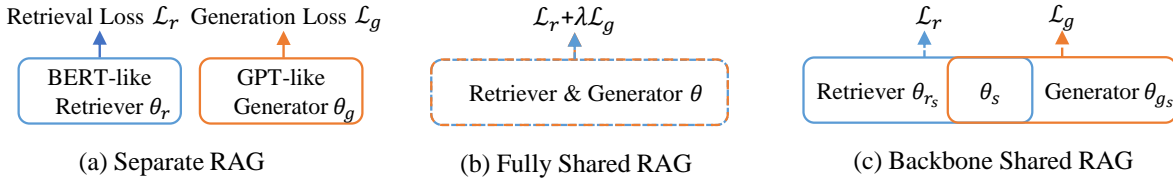**(a) Separate RAG**  **(b) Fully Shared RAG**  **(c) Backbone Shared RAG**

Figure 1: Comparing three categories of possible RAG frameworks: (a) most previous works lie in the separate RAG category, in which the retrieval task and the generation task cannot benefit from each other; (b) only a few trials are about the fully shared RAG, which may suffer from performance decrease due to negative transfer and require effort-taking loss balancing to determine $\lambda$; (c) what we proposed is a backbone shared RAG, which ensures effective knowledge transfer between the two tasks without the need to perform effort-taking loss balancing.

in the retrieval task and the generation task. Moreover, balancing the retrieval loss and the generation loss is non-trivial and may lead to an effort-taking hyperparameter search.

To address the two challenges, we first propose the WorthBuying dataset with 735K high-quality documents, 50K Question-Document-Answer (QDA) tuples, and human annotated test data of relevant documents for 1K questions and 500 QA pairs (*C1*). The knowledge base in our dataset comes from professional users, reducing conflicts and errors, and is more informative, with 1.1K words per document rather than a few dozen words as in existing e-commerce knowledge bases. We also annotate high-quality QA pairs with GPT-4 (Achiam et al., 2023) and manually review the test set. We then propose a specifically designed BSharedRAG framework that effectively adapts RAG to the e-commerce domain (*C2*). As shown in Figure 1 (c) we apply two task-specific modules to independently minimize the retrieval loss and generation loss, which avoids effort-taking loss balancing. With the shared backbone that benefits from domain-specific continual pre-training, the retriever and the generator can benefit from each other to improve both retrieval and generation performance.

Our contributions are summarized as follows:

• We construct a high-quality dataset called **WorthBuying** for the e-commerce domain, which contains 735K documents, 50K QDA tuples for training and some high-quality human labeled test data to facilitate further research.

• We propose a backbone-shared RAG framework (**BSharedRAG**) for the e-commerce domain, which enables effective knowledge transfer between the retrieval task and the generation task to improve the performance of both tasks. Although we focus on e-commerce, the framework can be generalized to other domain-specific RAG models.

• Experimental results demonstrate that our BSharedRAG, with affordable efficiency, outperforms traditional RAG methods in both retrieval and generation evaluation.

## 2 Related Work

**LLMs for the E-Commerce Domain.** To adapt LLMs for the e-commerce domain, some studies leverage LLMs like GPT-3.5-turbo and GPT-4 to generate e-commerce instruction tuning datasets (Li et al., 2024; Peng et al., 2024). Continual pre-training on e-commerce data has been shown to further enhance LLM performance in e-commerce tasks (Ma et al., 2023a). Despite these advancements, issues such as hallucination hindering real-world application, especially for Product Question Answering (PQA) tasks that heavily rely on product knowledge (Miller et al., 2020; Shi et al., 2023a; Li et al., 2024). Existing PQA datasets only contain short, non-informative knowledge bases about products, and QA pairs are of various quality (Gupta et al., 2019a; Gao et al., 2019a; Shen et al., 2023a; Deng et al., 2023; Chen et al., 2019a), since they contain subjective or conflicting answers provided by users. In this paper, we address the aforementioned issues by proposing a high-quality WorthBuying dataset. Moreover, we propose a BSharedRAG framework to improve retrieval and generation performance significantly.

**Retrieval Augmented Generation for LLMs.** Retrieval Augmented Generation (RAG), where the retriever provides relevant knowledge from external sources and LLMs answer the query based on the retrieved results, significantly reduces problems such as hallucinations and untimely knowledge updating of LLMs (Zhu et al., 2024; Zhao et al., 2024; Gao et al., 2024). In previous SeperateRAG frameworks, the retriever and the generator are typically separate and share no parame-
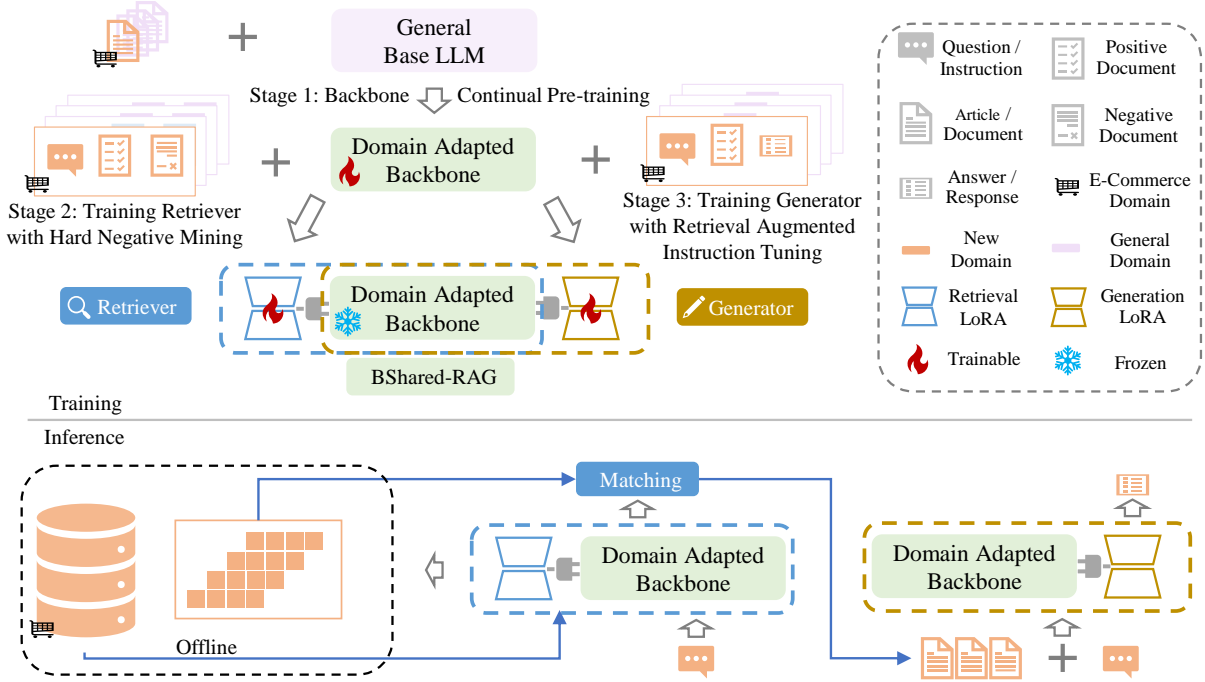
Figure 2: Overview of training and inference of our proposed BSharedRAG Framework.

ters, with the retrievers being off-the-shelf embedding models like DPR (Karpukhin et al., 2020), BGE (Xiao et al., 2023), E5 (Wang et al., 2022), or BERT-like architectures (Devlin et al., 2019), and the generators being LLMs (Touvron et al., 2023a,b; Yang et al., 2023). Some RAG frameworks, such as RETRO (Borgeaud et al., 2022), REPLUG (Shi et al., 2023b), and RA-DIT (Lin et al., 2023), make some performance improvements within the SeperateRAG framework. These frameworks require an independent retriever alongside the language model, without any shared parameters between the retriever and generator, creating a gap between these components. It is challenging for them to share the benefits derived from continual pre-training (CPT), making it difficult for them hard to adapt to a specific domain. Recently, GritLM (Muennighoff et al., 2024) is proposed, in which the retriever and generator share all parameters and thus the the retriever can benefit from the capacity of LLM. However, it is unclear how GritLM can be adapted to domain-specific scenarios: it requires large-scale training data and high computation costs that can hardly be satisfied in a specific domain. In comparison, we propose a lightweight framework that can efficiently and effectively convert a continually pre-trained domain-specific LLM to a backbone-shared domain-specific RAG model.

# 3 Method

In this section, we introduce our BSharedRAG for scenarios that require extensive domain knowledge, such as e-commerce.

## 3.1 Framework

**Separate RAG.** Traditional RAG frameworks treat retrieval and generation as two distinct tasks, where the retriever and generator share no parameters (Figure 1(a)). Formally, they optimize two groups of parameters separately:

$$\min \mathcal{L}_r(\theta_r), \quad \min \mathcal{L}_g(\theta_g) \quad (1)$$

where $\theta_r$ and $\mathcal{L}_r$ are the parameters and loss for retrieval model, and $\theta_g$ and $L_g$ are those for generation model. This paradigm cannot benefit from the shared information between the two related tasks as discussed in Section 1, which limits the performance of both retrieval and generation.

**Fully Shared RAG.** One straightforward method to overcome the problems is to build a shared model that can be used for retrieval and generation based on multi-task learning (Figure 1 (b)):

$$\min \mathcal{L}_r(\theta) + \lambda \mathcal{L}_g(\theta) \quad (2)$$

where $\theta$ is the parameters of the shared model, and $\lambda > 0$ is the weight for balancing the two losses. This framework may suffer from negative transfer,

i.e., performance decrease in both tasks due to the potential conflicts between them, as verified in our experiments. Moreover, finding an appropriate $\lambda$ to ensure loss balancing is non-trivial and may require an effort-taking hyperparameter search.

**Backbone Shared RAG.** To make both retrieval and generation benefit from the capability of LLM, and avoid the difficulty of balancing two objectives, we propose a BSharedRAG framework. The basic idea is to use a frozen LLM backbone as a shared information processor and two task-specific modules for retrieval and generation tasks.We use LoRA (Hu et al., 2021) method to finish this idea. Under this design, the two tasks can share knowledge from the backbone, *e.g.*, domain-specific continual pre-training, while using task-specific parameters for different purposes to avoid the difficulty of balancing two tasks and negative transfer:

$$\min \mathcal{L}_r(\theta_{r_s}; \theta_s), \quad \min \mathcal{L}_g(\theta_{g_s}; \theta_s) \qquad (3)$$

where $\theta_{r_s}$ and $\theta_{r_s}$ are task-specific modules upon a pre-trained shared backbone model $\theta_s$.

### 3.2 Training Strategies

As shown in Figure 2, we adopt the following three training stages to optimize Equation (3).

**Stage 1: Backbone Continual Pre-training.** We leverage continual pre-training to adapt a general domain LLM (Baichuan2-7B-Base (Yang et al., 2023) here) to a specific domain such as e-commerce (here called Ecom-base). The domain-specific continual pre-training objective is the next token prediction task:

$$\max_{\theta_s} \sum_{i=1}^{N} \log p(y_i | y_{<i}; \theta_s) \qquad (4)$$

For the training data, we mix corpus from the e-commerce domain and the general domain to avoid catastrophic forgetting following (Xie et al., 2023; Ke et al., 2023; Que et al., 2024).

**Stage 2: Training Retriever with Hard Negative Mining.** We then minimize $\mathcal{L}_r$ by fine-tuning a domain-specific retrieval model based on a frozen base LLM $\theta_s$ with trainable LoRA (Hu et al., 2021) module $\theta_{r_s}$. As our retrieval model adopts a GPT-like architecture, we adhere to the RepLLaMA paradigm (Ma et al., 2023b) to derive sentence embeddings from the End of Sentence (EOS) token.

Following the approach of the bi-encoder dense retriever DPR (Karpukhin et al., 2020), we employ in-batch contrastive learning to fine-tune Ecom-base with InfoNCE loss:

$$\mathcal{L}_r = -\log \frac{e^{sim(q,D^+)}}{e^{sim(q,D^+)} + \sum_{D_i^- \in D^-} e^{sim(q,D_i^-)}}$$

$$\qquad (5)$$

Here, $D^+$ represents the relevant document, and $D^-$ denotes a set of irrelevant documents to the query. We employ the in-batch negatives method, where one positive document and several negative documents are included in one sample.

We also propose a strategy for identifying hard negatives during contrastive learning. Specifically, we utilize a baseline retrieval model to retrieve a preliminary set of potential negatives, then the high-ranking yet erroneously matched documents, *i.e.*, hard negatives, which helps refine the model's discrimination capabilities.

For our hard-negative mining, we first employ a base retrieval model that is fine-tuned on random negative samples to rank documents. Then we utilize the top 30 but incorrect results as hard-negative samples. Experimental results show that this strategy enhances the effectiveness of retrieval.LoRA is applied to all layers of our LLMs. The targeted weights for LoRA adaptation consist of the query, key, value, and output weights $(W_q, W_k, W_v, W_o)$ within the attention modules, as well as the weights in the multi-layer perceptron (MLP) components.

**Stage 3: Training Generator with Retrieval Augmented Instruction Tuning.** To ensure that the generator can effectively utilize retrieved information, we augment the instruction-response pairs by incorporating the retrieved articles, forming triples $(q, d, a)$, where $q$ denotes question tokens, $d$ denotes the retrieved document, and $a$ denotes the answer. Specifically,

$$\mathcal{L}_g = -\sum_{i=1}^{N} \log p(a_i | a_{<i}, q, d, \theta) \qquad (6)$$

Similar to the embedding model, we employ the LoRA module to fine-tune only a subset of additional parameters $\theta_{g_s}$ on top of foundational LLM.

### 3.3 Inference

During inference, the retriever is first employed to obtain documents, which are subsequently given to the generator for processing (Figure 2). Specifically, we first prepare all embeddings for the documents in the domain-specific knowledge base by

| Dataset | Document Source | # Categories | # Document | Avg words per Doc | Release |
|---|---|---|---|---|---|
| AmazonQA (Gupta et al., 2019b) | PR | 17 | 923K | 63.41 | ✓ |
| SubjQA (Bjerva et al., 2020) | PR | 6 | 10,098 | 289.87 | ✓ |
| semiPQA (Shen et al., 2022) | PI | - | 11,243 | - | ✗ |
| xPQA (Shen et al., 2023b) | PR | - | 2,500 | 76.87 | ✓ |
| JD (Gao et al., 2019b) | PR & PI | 38 | 469,955 | 16.94 | ✓ |
| Taobao (Chen et al., 2019b) | PR | 2 | 1,155,530 | 74.15 | ✗ |
| WorthyBuying (ours) | PA | 25 & 128 | 735,937 | **1171.25** | ✓ |

Table 1: Comparison of Product Question Answering (PQA) datasets. The document types are classified into Product Reviews (PR), Product Information (PI), and **Product Analysis from professional users (PA)** .

using the BSharedRAG retriever offline. In the on-line stage, we only extract the embedding of the question and use it to retrieve the most relevant $N$ documents ($N = 3$). Finally, we use the BShare-dRAG generator to obtain an answer to the question by taking the top documents as input.

## 4   WorthBuying Dataset

Existing e-commerce datasets are mainly constructed from user-generated product reviews, which are brief and noisy, resulting in limited assistance to LLMs in generating rich content (Wan and McAuley, 2016; Gupta et al., 2019a). To solve this issue, we build a WorthyBuying dataset, which includes an accurate knowledge base and high-quality QDA tuples for the e-commerce domain.

**Data Collection.**   To obtain a high-quality corpus for the e-commerce domain, we collect 1.1M professional product reviews from an e-commerce website[1]. Beginning from the raw corpus, we conduct several steps of data cleaning to ensure the quality, such as filtering contents with dirty words[2], removing any HTML tags and images, and deleting those with the length beneath a threshold. We then adopt the Minihash method[3] to eliminate duplicate entries, which finally results in 735k cleaned documents. More details are presented in the Appendix.

**QDA Tuples Generation.**   Users often ask LLMs to obtain information about products in the e-commerce domain. However, the questions and answers in existing datasets are often derived from brief user QA comments, which are neither verified nor grounded in any relevant documents. The informative review documents we collected can provide sufficient clues to answer product-related ques-

tions. To facilitate documents lacking high-quality QDA tuples aligned with user needs, we leverage GPT-4 to generate questions and answers by prompt with these documents. We collect 50K such ⟨question, document, answer⟩ tuples grounded on documents. This dataset can support the training of the embedding model and generation model.

**Dataset Statistics and Comparisons.**   As indicated in Table 1, previous datasets consist of undetailed short documents. Their questions and answers are directly collected from brief user QA comments without verification, leading to noise and even incorrect answers. Our WorthBuying dataset contains much longer documents with more than 50k QDA tuples grounded on 735k informative documents. We also contain the most detailed product categories than existing datasets.

## 5   Experiment

### 5.1   Experiment Setup

#### 5.1.1   Datasets and Evaluation Metrics

To evaluate the retrieval effectiveness, we use an existing dataset **CPR-Ecom** (Long et al., 2022), which contains a corpus of 1,002,822 Chinese documents, 100k human annotated query and relevant document pairs for training and 1k for testing in an e-commerce domain, and our constructed **WorthBuying** dataset, which contains a corpus of 735k Chinese documents, 50k GPT-4 generated questions from 50k documents for training and 1k GPT-4 generated questions with 10.2k human-annotated documents with 1 or 2 relevance ratings for testing, as benchmarks. We adopt two retrieval metrics for evaluating our models: nDCG (Normalized Discounted Cumulative Gain) and Hit Rate. nDCG@n measures the ranking quality of top n retrieved documents, considering accumulated gains from relevant documents while discounting their positions. Hit@n assesses the presence of relevant documents

---

[1]https://www.smzdm.com/
[2]https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words.
[3]https://github.com/ChenghaoMou/text-dedup

| Model | CPR-Ecom | | | | WorthBuying | | | |
|---|---|---|---|---|---|---|---|---|
| | nDCG@3 | Hit@3 | nDCG@5 | Hit@5 | nDCG@3 | Hit@3 | nDCG@5 | Hit@5 |
| Multi-E5-large (Wang et al., 2024b) | 42.44 | 49.20 | 45.36 | 56.30 | 34.98 | 39.50 | 37.51 | 45.20 |
| M3E-large (Wang Yuxin, 2023) | 44.89 | 51.70 | 47.65 | 58.40 | 43.02 | 49.40 | 45.64 | 55.40 |
| BGE-large-zh (Xiao et al., 2023) | 51.25 | 58.60 | 54.44 | 66.30 | 47.01 | 53.50 | 49.86 | 56.30 |
| BGE-large-zh + HN | 52.38 | 60.35 | 56.23 | 68.50 | 48.46 | 55.70 | 50.83 | 58.30 |
| BGE-large-zh + CPT + HN | 47.15 | 58.10 | 50.28 | 61.20 | 45.38 | 51.20 | 46.36 | 55.40 |
| FullyShared-RAG | 38.67 | 44.80 | 43.24 | 48.80 | 33.28 | 36.70 | 36.56 | 41.20 |
| **BShared-RAG Retriever (ours)** | **61.39** | **68.40** | **64.28** | **75.40** | **51.83** | **58.60** | **54.06** | **63.90** |
| | (+17%) | (+13%) | (+14%) | (+10%) | (+7%) | (+5%) | (+6%) | (+10%) |
| w/o CPT | 56.74 | 63.80 | 58.80 | 67.90 | 49.02 | 54.50 | 51.78 | 58.30 |
| w/o HN | 58.35 | 64.70 | 61.23 | 70.60 | 49.34 | 56.30 | 52.03 | 60.90 |
| w/o CPT & HN | 53.84 | 61.50 | 57.76 | 66.30 | 47.24 | 53.40 | 49.93 | 57.20 |

Table 2: Comparing retrievers of different RAG frameworks. CPT denotes continual pre-training and HN denotes using hard negative samples. Our BShared-RAG Retriever outperforms all baselines by a large margin. CPT fails to help the BGE adapt to the e-commerce domain and even hurts the performance. FullShared-RAG performs the worst, showing that sharing all parameters between retrieval and generation leads to severe performance degradation.

within the top n results.

To evaluate generation quality, we use our built WorthBuying dataset. We randomly sample 500 QA-pairs from the test set and hire human annotators to verify and correct the questions and answers. This manual correction ensures the accuracy of the results and provides a more reliable evaluation. We employ a comprehensive set of widely used automatic metrics for evaluation: n-gram based BLEU-3 (Papineni et al., 2002) and ROUGE-L (Lin, 2004), BERTScore (Zhang et al., 2020) which calculates semantic similarity of ground-truth and generated answers based on a BERT model. Accuracy (Es et al., 2023) is calculated by GPT-4 to evaluate question and answer pairs. More details are in the Appendix C.

### 5.1.2 Baselines

We compare our BSharedRAG Retriever with several widely-used baseline retrieval models:

**BGE** (Xiao et al., 2023) is the state-of-the-art BERT-like model trained with multi-stage contrastive learning using over 200M paired data. To ensure a fair comparison, we fine-tune BGE through continual pre-training and in-domain contrastive learning with hard negative samples, resulting in two variants.

**M3E** (Wang Yuxin, 2023) is an open-source text embedding model trained by a massive corpus with over 22M sentence pairs, supporting bilingual (Chinese-English) text retrieval.

**Multilingual-E5** (Wang et al., 2024b) transfers E5 embedding model (English only) (Wang et al., 2022) to 100+ languages with a two-stage train-ing and exhibits competitive performance across a broad range of languages.

For generation, we choose the following representative LLMs from both open domains and e-commerce domain as baselines for comparison:

**GPT-3.5** and **GPT-4** (Achiam et al., 2023) are developed by OpenAI which are the most advanced LLMs. We feed them with retrieved passages in the prompt for a fair comparison.

**EcomGPT** (Li et al., 2024) is an e-commmerce domain-specific LLM trained on BLOOMZ-7B (Muennighoff et al., 2022) with their close-source instruction dataset EcomInstruct.

**Baichuan2-7B-Chat** (Yang et al., 2023) is one of the best open-source Chinese LLMs. It adopts 100k samples for supervised fine-tuning, after which is optimized using Reinforcement Learning from Human Feedback (RLHF) (Ziegler et al., 2020).

**Seperate RAG** has separate parts for retrieval and generation tasks, as shown in 1(a). To make sure the same parameters as BSharedRAG in two stages, we train Llama3-8b (Dubey et al., 2024), a powerful foundation LLM and eCeLLM-M(7B) (Peng et al., 2024), a powerful e-commerce domain LLM with the same data, and configuration as BSharedRAG-retriever in retrieval stage. For the generation, we use the BSharedRAG-generator.

**Fully Shared RAG** integrates retrieval and generation into a single model, as shown in 1(b). Using the same base model, data, and configuration as BSharedRAG, it employs a single LoRA to fully share retrieval and generation parameters.

| Model | BLEU-3 | ROUGE-L | BERTScore | Accuracy (%) |
|---|---|---|---|---|
| *FullySharedRAG as retriever* | | | | |
| FullySharedRAG | 2.35 | 2.58 | 62.37 | 68.34 |
| *SeperateRAG as retriever* | | | | |
| SeperateRAG (llama3 retriever) | 10.35 | 10.38 | 71.84 | 81.38 |
| SeperateRAG (ecellm retriever) | 10.13 | 11.23 | 71.98 | 82.41 |
| *BSharedRAG Retriever as retriever* | | | | |
| GPT-3.5 | 16.07 | 12.76 | 74.33 | 86.77 |
| GPT-4 (Achiam et al., 2023) | 19.76 | 15.92 | 77.98 | 89.86 |
| EcomGPT (Li et al., 2024) | 6.91 | 8.12 | 69.18 | 78.70 |
| Baichuan2-7b-base | 2.32 | 2.49 | 62.56 | 76.81 |
| Baichuan2-7b-base + CPT | 2.46 | 3.24 | 62.58 | 80.85 |
| Baichuan2-7b-base + RAG-IT | 9.24 | 9.83 | 70.07 | 82.91 |
| Baichuan2-7b-chat (Yang et al., 2023) | 8.82 | 11.21 | 72.45 | 79.84 |
| Baichuan2-7b-chat + RAG-IT | 10.24 | 12.34 | 72.12 | 82.33 |
| **BSharedRAG Generator (ours)** | **12.63 (+23%)** | **12.85 (+4%)** | **74.75 (+3%)** | **84.26 (+2%)** |

Table 3: Evaluation of generation results based on different retrievers on the WorthBuying-PQA test set. RAG-IT denotes retrieval augmented instruction tuning. The FullySharedRAG method performs worse because the generation objective may conflict with the retrieval objective. Compared with Baichuan2-7b series of baselines, our model achieves the best performance, demonstrating both CPT and RAG-IT contribute to the final performance.

## 5.2 Experimental Results

### 5.2.1 Retrieval Evaluation

Results in Table 2 indicate that our *BShared-RAG Retriever* significantly outperforms all baselines in all metrics upon both datasets. *BGE-large-zh+HN*, which is fine-tuned on our dataset using hard negative contrastive learning, serves as the strongest baseline. However, our model still achieves improvements ranging from 5% to 17% over this baseline. The *FullyShared-RAG* performs much worse than ours and most baselines in separate RAGs. This indicates that it is difficult to optimize two objectives with fully shared parameters. Compared to BERT-like retrievers, *e.g.*, BGE-large-zh, our proposed method can effectively benefit from continual pre-training. In contrast, *BGE-large-zh+CPT+HN* has significant drops in all metrics when applying continual pre-training. In our analysis, there is evidence indicating that embedding models based on architectures like BGE require more fine-tuning training data after CPT, potentially reaching the terabyte (TB) level (Wang et al., 2024a). Fine-tuning with the same amount of data as BSharedRAG is insufficient for the BGE model, which could lead to a decline in performance. We conduct ablation studies to analyze the detailed contributions of each module. Both CPT and HN contribute to significant and consistent improvements, while CPT brings more gains.
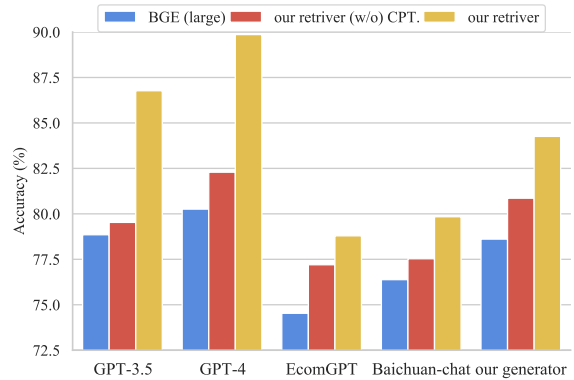


Figure 3: Evaluating the influence of different retrievers to generation effectiveness. CPT is continual pre-training, which benefits retrieval effectiveness a lot via sharing an LLM backbone. Accuracy is judged by GPT-4. Other metrics are not shown due to limited space, but we observe similar trends.

### 5.2.2 Generation Evaluation

Results in Table 3 show that our model performs the best among open-source models. In terms of ROUGE-L and BertScore, our model is even slightly better than GPT-3.5. FullyShared-RAG performs the worst because it cannot well balance the retrieval and generation tasks. Baichuan2-7b-base has no instruction following abilities and thus it performs lowest in QA testing. After RAG-IT, it can obtain instruction following abilities and thus its performance is significantly improved, *e.g.*, from 76.81% to 82.91% in Accuracy. A strong chat model, such as Baichuan2-7b-chat, can be also improved by RAG-IT because it brings domain-

| Retrieval Model | Kendall's $\tau$ | Correlation |
|---|---|---|
| BGE-large-zh | 0.0498 | 0.045 |
| Llama3-retriever | 0.0937 | 0.085 |
| Ecellm-retriever | 0.1032 | 0.119 |
| BSharedRAG Retriever | 0.1433 | 0.155 |

Table 4: Measuring whether a retriever favors a sentence, which is easy to generate from a prompt of a question. We calculate Kendall's tau and Pearson's correlation between two ranked lists, one by a retrieval model and the other by the generation probability estimated by BSharedRAG Generator. The larger the better.

specific knowledge. Our proposed BShared-RAG generator performs better than *Baichuan-7b-base + RAG-IT* by 37% in BLEU-3 and 31% in ROUGE-L. These results demonstrate CPT and RAG-IT are both necessary and effective to improve generators. Meanwhile, the Separate-RAG model, which employs Llama3 and EcellM as retrievers, maintains the same parameters as BSharedRAG but exhibits lower performance. This result underscores the importance of sharing a unified backbone, confirming its necessity for enhanced performance.

### 5.2.3 Influence of Retrieval to Generation

We further conduct some experiments to analyze how retrievers influence generation. We compare the accuracy of generated results by models using BGE-large-zh, our BShared-RAG retriever and BShared-RAG w/o CPT. Results in Figure 3 indicate that our BShared-RAG Retriever can significantly improve generation accuracy no matter using GPT-3.5/GPT-4 or open-source models. Without CPT, the performance has a large drop. This underscores how important it is to bring continual pre-training into retrieval models.

### 5.3 Analysis of BSharedRAG Performance Gain

Reasons for BSharedRAG's outstanding performance can be attributed to following three factors:

**BSharedRAG provides a way to scale up the retriever parameters appropriately**. Conventional BERT-based embedding models (Devlin et al., 2019; Wang et al., 2022; Xiao et al., 2023) face inherent architectural limitations that hinder their ability to scale up parameters, thus preventing them from fully leveraging the benefits of scaling laws. While some approaches have employed LLM as retrievers, these methods often consume excessive memory. BSharedRAG effectively addresses

these challenges by sharing a common backbone between the retriever and generator, allowing the retriever to benefit from larger parameter sizes without increasing the overall computational overhead.

**BSharedRAG enables both the retriever and generator to benefit from continued pre-training**. In SeparateRAG, the retriever and generator are trained independently, meaning the benefits of continued pre-training cannot be shared between them. On the other hand, in the FullySharedRAG structure, the retrieval loss and generation loss can interfere with each other, leading to degraded performance. In contrast, the BSharedRAG framework requires only continued pre-training (CPT) on the backbone LLM, allowing both the retriever LoRA and generator LoRA to benefit from domain-specific performance improvements achieved through CPT.

**BSharedRAG facilitates better preference alignment between the retriever and generator**. To verify this, we conducted an experiment. For each question, we gathered the ground-truth document title and top non-ground-truth titles returned by two methods. We then concatenated each question with a document title and calculated the generation probability using our BSharedRAG Generator. Next, we computed Kendall's tau and Pearson correlation between ranked lists from either BGE-large-zh or our retriever, and from the generator. As shown in Table 4, our method achieves significantly higher correlations, indicating that the BSharedRAG Retriever favors texts with higher generation probabilities, helping it outperform separate RAG baselines.

As illustrated in Figure 4, BGE-large-zh prioritizes titles closely matching the question, whereas our BShared-RAG Retriever ranks titles that better answer the question. For example, our method ranks "Equipped with a new 33 million pixel sensor, Sony officially released the full-frame mirrorless A7M4" higher, which is like a more relevant answer to the question. In contrast, BGE-large-zh ranks "Sony A7M4 after buying user experience" without "pixel" information higher, or another document the second, which is hard for the generator to capture due to its isolation in the middle (Liu et al., 2024).

### 5.4 Inference space cost and time

We evaluate the inference space cost and time of our BSharedRAG framework. As demonstrated in Table 5, BSharedRAG has a lower space cost than

**Question:** How high are the **pixels** of the **Sony A7M4 camera** and what is **user experience** like?

**Top Documents by BSharedRAG Retriever**

{'title': '**Equipped with a new 33 million pixel sensor, Sony officially released the full-frame mirrorless A7M4**', 'contents': '... Sony today officially announced the launch of the A7M4, which uses a new 33 million effective pixel back-illuminated sensor, as well as ...'}

{'title': 'Sony A7M4 | 2022's most balanced and comprehensive mirrorless camera', 'contents': 'As an old user of A7M3, I waited for 3 years and finally waited for the upgraded A7M4, I grabbed this camera from the Sony Mall in November last year, **after 4 months of in-depth use, I am still very satisfied, as an all-round balanced bucket machine, both the quality of the photo and the quality of**...'}

{...}  Easy to use in generation

**Top Documents by BGE-large-zh**

{'title': '**Sony A7M4 after buying user experience**', 'contents': 'Due to the increase in business volume this year, the original camera is not enough to cope with daily shooting. So I started planning to change the camera in March. I've been hesitating between Canon and Sony... '}

No information about pixels

{'title': 'Is it worth buying the **Sony A7M4** that thousands of people snapped up', 'contents': 'Today I'll talk to you about Sony 7M4 . ... Main parameters: full-frame approx. **33 million pixels**, back-illuminated CMOS 5-axis body image stabilization, BIONZ XR image processor ...'}

{...}  Not used in generation

**Answer by BSharedRAG Generator**

Sony officially released the full-frame mirrorless A7M4, which has a whopping 33 million effective pixels, compared to the previous A7M2, it has a significant increase in pixels, . ... Sony A7M4 is a powerful and easy-to-use camera that offers a wealth of features and excellent image quality.

**Answer by BSharedRAG Generator**

Sorry, from the above reference information, I don't know how many pixels the Sony A7M4 camera has. But the Sony A7M4 is an excellent camera with excellent performance and features for those who are looking for high image quality and performance.
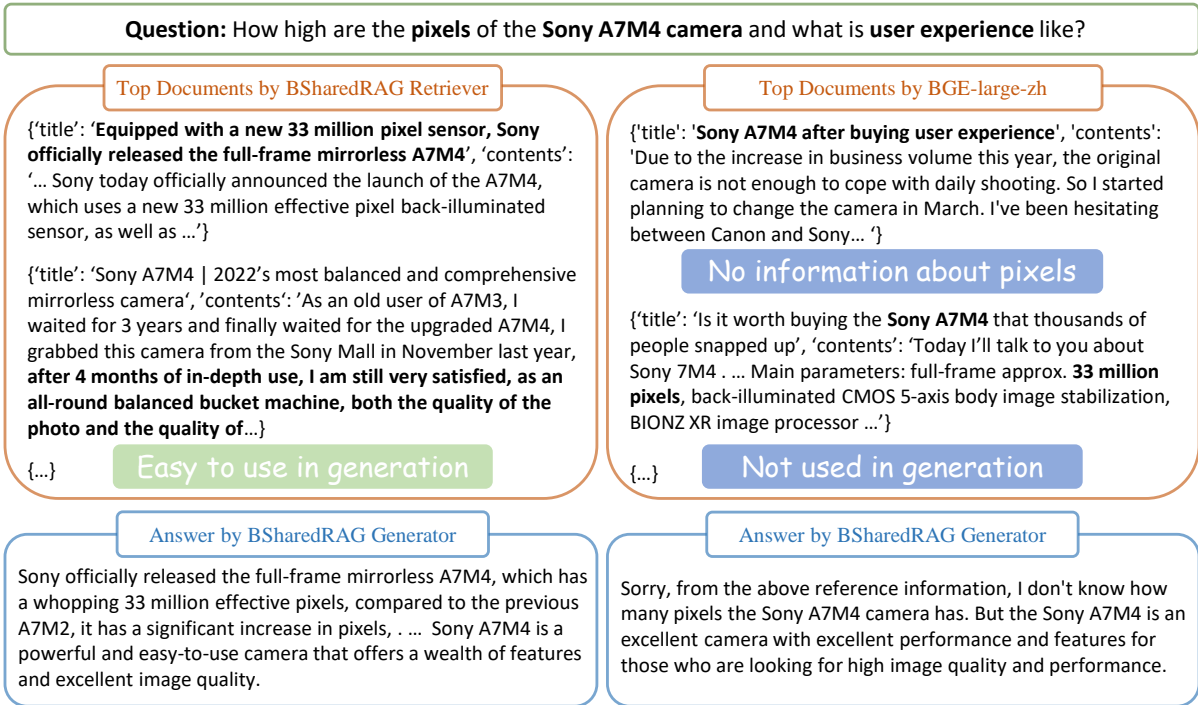
Figure 4: A representative example to compare our BSharedRAG with a separate RAG. For the given question, our BSharedRAG Retriever favors the documents, in which some sentences are easy to be generated from the prompt of question. In contrast, the BERT-like BGE-large-zh model tends to retrieve some documents, in which some sentences match the question well. However, such document may be less suitable for generating answers due to some issues, *e.g.*, important information missing or not easy to be used by generators.

| Model | Parameters | RAM | VRAM | Storage |
|---|---|---|---|---|
| SeperateRAG (bge-zh-large) | 7.87B | 30.07GB | 58.63GB | 15.19GB |
| SeperateRAG (llama3-retriever) | 15.84B | 61.19GB | 120.38GB | 29.05GB |
| **BSharedRAG** | **7.54B** | **29.02GB** | **56.61GB** | **14.09 GB** |

Table 5: Comparing the space costs of different RAG frameworks, SeperateRAG (bge-zh-large) and Seperat-eRAG (llama3-retriever) respectively utilize the bge-zh-large and finetuned llama3 models, as retrievers, while sharing the same generator as BSharedRAG.

| Framework | Average Inference Time |
|---|---|
| SeperateRAG (bge-zh-large) | 0.65s |
| BSharedRAG | 0.91s |

Table 6: We calculate the average inference time over 1000 questions randomly sampled from WorthyBuying Dataset. Both BSharedRAG retriever and generator are accelerated by vllm (Kwon et al., 2023)
.

SeperateRAG due to eliminating the need for an additional retrieval component. In the retrieval part, BSharedRAG integrates a shared backbone model, which results in a marginal increase of retriever parameters compared to the BERT-based model. However, this does not lead to a significantly longer inference time. As shown in Table 6, the overall time increase is just 0.34 seconds, which remains within an acceptable range.

## 6 Conclusion

In this paper, we construct a high-quality e-commerce dataset called WorthBuying, which contains an e-commerce knowledge base and QA pairs for training and evaluating retrieval and generation

effectiveness. We also design BSharedRAG, which can effectively adapt RAG models to a specific domain. In our framework, we propose sharing the backbone model that benefits from domain-specific continual pre-training and using hard negative mining and RAG instruction tuning to optimize two Low-rank Adaptation (LoRA) modules in the retriever and generator respectively. Compared to the fully shared RAG framework, our method can avoid the issue of negative transfer between tasks and the difficulty of loss balancing. Experimental results demonstrate that compared to existing separate RAG frameworks, both the retrieval performance and the generation performance have been significantly improved.

## Limitations

The limitations of this work are as follows: (1) Due to the limited budget, we currently use only Chinese models and datasets in e-commerce domain in our experiments. As the proposed method is language and domain independent, it shall be applicable to more. Thus we plan to extend our methods to more languages and domains. (2) We implement a fully shared RAG framework based on one LoRA in this work. We are interested to explore more possibilities to compare two frameworks. (3) Our approach focuses on a basic retrieval and generation architecture. In the future, we plan to explore more advanced methodologies, such as after-retrieval Chain-of-Thought, re-ranking, and query rewriting, to enhance RAG's capability and robustness in the e-commerce domain.

## Ethical Statement

Our work aims to adapt a general LLM to the E-commerce domain, but the models we train may have negative impacts. For example, they could be used inappropriately, although we have performed data cleaning to avoid offensive content. However, this is a common issue currently faced in the LLM field, and it is not amplified by this work. In the future, we will consider more work on the safety of LLMs to optimize their security in the E-commerce domain. To protect the intellectual property rights of the data, we will strictly limit the dissemination of this dataset to academic research purposes only.

## References

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

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Johannes Bjerva, Nikita Bhutani, Behzad Golshan, Wang-Chiew Tan, and Isabelle Augenstein. 2020. Subjqa: A dataset for subjectivity and review comprehension. In *EMNLP 2020*, pages 5480–5494.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

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. *Preprint*, arXiv:2005.14165.

Jie Chen, Zhipeng Chen, Jiapeng Wang, Kun Zhou, Yutao Zhu, Jinhao Jiang, Yingqian Min, Wayne Xin Zhao, Zhicheng Dou, Jiaxin Mao, et al. 2024. Towards effective and efficient continual pre-training of large language models. *arXiv preprint arXiv:2407.18743*.

Shiqian Chen, Chenliang Li, Feng Ji, Wei Zhou, and Haiqing Chen. 2019a. Review-driven answer generation for product-related questions in e-commerce. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 411–419, New York, NY, USA. Association for Computing Machinery.

Shiqian Chen, Chenliang Li, Feng Ji, Wei Zhou, and Haiqing Chen. 2019b. Review-driven answer generation for product-related questions in e-commerce. In *WSDM 2019*, pages 411–419.

Yang Deng, Wenxuan Zhang, Qian Yu, and Wai Lam. 2023. Product question answering in E-commerce: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11951–11964, Toronto, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019*, pages 4171–4186.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation. *Preprint*, arXiv:2309.15217.

Shen Gao, Zhaochun Ren, Yihong Zhao, Dongyan Zhao, Dawei Yin, and Rui Yan. 2019a. Product-aware answer generation in e-commerce question-answering. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 429–437, New York, NY, USA. Association for Computing Machinery.

Shen Gao, Zhaochun Ren, Yihong Zhao, Dongyan Zhao, Dawei Yin, and Rui Yan. 2019b. Product-aware answer generation in e-commerce question-answering. In *WSDM 2019*, pages 429–437.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Mansi Gupta, Nitish Kulkarni, Raghuveer Chanda, Anirudha Rayasam, and Zachary C. Lipton. 2019a. Amazonqa: A review-based question answering task. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4996–5002. International Joint Conferences on Artificial Intelligence Organization.

Mansi Gupta, Nitish Kulkarni, Raghuveer Chanda, Anirudha Rayasam, and Zachary C. Lipton. 2019b. Amazonqa: A review-based question answering task. In *IJCAI 2019*, pages 4996–5002.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2023. Continual pre-training of language models. *Preprint*, arXiv:2302.03241.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

langchain. 2023. langchain.

Yangning Li, Shirong Ma, Xiaobin Wang, Shen Huang, Chengyue Jiang, Hai-Tao Zheng, Pengjun Xie, Fei Huang, and Yong Jiang. 2024. Ecomgpt: Instruction-tuning large language models with chain-of-task tasks for e-commerce. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18582–18590.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, Luke Zettlemoyer, and Scott Yih. 2023. Radit: Retrieval-augmented dual instruction tuning. *Preprint*, arXiv:2310.01352.

Jerry Liu. 2022. LlamaIndex.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Dingkun Long, Qiong Gao, Kuan Zou, Guangwei Xu, Pengjun Xie, Ruijie Guo, Jian Xu, Guanjun Jiang, Luxi Xing, and Ping Yang. 2022. Multi-cpr: A multi domain chinese dataset for passage retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3046–3056.

Shirong Ma, Shen Huang, Shulin Huang, Xiaobin Wang, Yangning Li, Hai-Tao Zheng, Pengjun Xie, Fei Huang, and Yong Jiang. 2023a. Ecomgpt-ct: Continual pre-training of e-commerce large language models with semi-structured data. *Preprint*, arXiv:2312.15696.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2023b. Fine-tuning llama for multi-stage text retrieval. *Preprint*, arXiv:2310.08319.

John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. 2020. The effect of natural distribution shift on question answering models. *Preprint*, arXiv:2004.14444.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. *Preprint*, arXiv:2402.09906.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL 2002*, pages 311–318.

Bo Peng, Xinyi Ling, Ziru Chen, Huan Sun, and Xia Ning. 2024. ecellm: Generalizing large language models for e-commerce from large-scale, high-quality instruction data. *Preprint*, arXiv:2402.08831.

Jiantao Qiu, Haijun Lv, Zhenjiang Jin, Rui Wang, Wenchang Ning, Jia Yu, ChaoBin Zhang, Zhenxiang Li, Pei Chu, Yuan Qu, Jin Shi, Lindong Lu, Runyu Peng, Zhiyuan Zeng, Huanze Tang, Zhikai Lei, Jiawei Hong, Keyu Chen, Zhaoye Fei, Ruiliang Xu, Wei Li, Zhongying Tu, Lin Dahua, Yu Qiao, Hang Yan, and Conghui He. 2024. Wanjuan-cc: A safe and high-quality open-sourced english webtext dataset. *Preprint*, arXiv:2402.19282.

Haoran Que, Jiaheng Liu, Ge Zhang, Chenchen Zhang, Xingwei Qu, Yinghao Ma, Feiyu Duan, Zhiqi Bai, Jiakai Wang, Yuanxing Zhang, Xu Tan, Jie Fu, Wenbo Su, Jiamang Wang, Lin Qu, and Bo Zheng. 2024. D-cpt law: Domain-specific continual pre-training scaling law for large language models. *Preprint*, arXiv:2406.01375.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Preprint*, arXiv:2302.00083.

Xiaoyu Shen, Akari Asai, Bill Byrne, and Adrià de Gispert. 2023a. xpqa: Cross-lingual product question answering across 12 languages. *arXiv preprint arXiv:2305.09249*.

Xiaoyu Shen, Akari Asai, Bill Byrne, and Adria De Gispert. 2023b. xPQA: Cross-lingual product question answering in 12 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 103–115, Toronto, Canada. Association for Computational Linguistics.

Xiaoyu Shen, Gianni Barlacchi, Marco Del Tredici, Weiwei Cheng, and Adrià Gispert. 2022. semiPQA: A study on product question answering over semi-structured data. In *Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 111–120.

Kaize Shi, Xueyao Sun, Dingxian Wang, Yinlin Fu, Guandong Xu, and Qing Li. 2023a. Llama-e: Empowering e-commerce authoring with multi-aspect instruction following. *Preprint*, arXiv:2308.04913.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023b. Replug: Retrieval-augmented black-box language models. *Preprint*, arXiv:2301.12652.

Xiao Shi, Zhengyuan Zhu, Zeyu Zhang, and Chengkai Li. 2023c. Hallucination mitigation in natural language generation from large-scale open-domain knowledge graphs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12506–12521, Singapore. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Mengting Wan and Julian J. McAuley. 2016. Modeling ambiguity, subjectivity, and diverging viewpoints in opinion question answering systems. In *ICDM 2016*, pages 489–498.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Improving text embeddings with large language models. *Preprint*, arXiv:2401.00368.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Xiaohua Wang, Yuliang Yan, Longtao Huang, Xiaoqing Zheng, and Xuanjing Huang. 2023. Hallucination detection for generative large language models by Bayesian sequential estimation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15361–15371, Singapore. Association for Computational Linguistics.

He sicheng Wang Yuxin, Sun Qingxuan. 2023. Moka massive mixed embedding.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. 2023. Efficient continual pre-training for building domain specific large language models. *Preprint*, arXiv:2311.08545.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024. Raft: Adapting language model to domain specific rag. *Preprint*, arXiv:2403.10131.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *ICLR 2020*.

Zhebin Zhang, Xinyu Zhang, Yuanhang Ren, Saijiang Shi, Meng Han, Yongkang Wu, Ruofei Lai, and Zhao Cao. 2023. Iag: Induction-augmented generation framework for answering reasoning questions. *Preprint*, arXiv:2311.18397.

Lingjun Zhao, Khanh Nguyen, and Hal Daumé III. 2023a. Hallucination detection for grounded instruction generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4044–4053, Singapore. Association for Computational Linguistics.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. 2024. Retrieval-augmented generation for ai-generated content: A survey. *Preprint*, arXiv:2402.19473.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023b. A survey of large language models. *Preprint*, arXiv:2303.18223.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *Preprint*, arXiv:2403.13372.

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. *Preprint*, arXiv:2308.07107.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences. *Preprint*, arXiv:1909.08593.

## A Details of WorthBuying Dataset

### A.1 Examples

We show two examples of our constructed WorthBuying Dataset.

---

**title**: Upgraded from Xiaomi's first robot vacuum to the Roborock G10S—here's my honest take.

**contents**:
In today's digital landscape, authentic product reviews are a rarity amidst the overwhelming amount of marketing content. Motivated by this, I've decided to share my genuine experience with the original Xiaomi and Roborock G10s vacuum robots to help potential buyers get a clearer picture.

Roborock is praised online for its advanced algorithms and reliable quality. Conversely, Ecovacs is often criticized for poor algorithms and quality, despite being the top offline seller. Xiaomi and other similar brands are noted for their cost-effectiveness but lack distinctive features. I purchased the first-generation Xiaomi robot in 2017 for ¥1699, and it's still running well without any major part replacements except for the filter. The Roborock G10s, bought during a sale in 2023 for ¥3999 without any freebies, was intended for my 101m² apartment with multiple rooms and balconies. Mapping issues occurred with the Roborock during the first two attempts, taking over 17 minutes each, but it succeeded on the third try in 11 minutes. Both robots struggle with certain thresholds and may require virtual barriers if doors are not closed. Cleaning efficiency seems stagnant over the years. The Roborock did not significantly outperform Xiaomi despite newer technology, spending 65 minutes to clean a slightly smaller area than Xiaomi's 60 minutes for a similar space. Obstacle avoidance is one area where Roborock excels, identifying and navigating around small objects like cables and tissues, which Xiaomi tends to run over.
The mopping function of the Roborock was adequate, though it left water streaks, and the overall cleaning process was lengthy, especially when using separate sweeping and mopping modes. The Roborock app is user-friendly, offering customizable cleaning modes and map management, although it occasionally fails to retrieve maps.

In summary, while Roborock recognizes more obstacles and has a better mopping system, the difference in everyday use is minimal. Both require preliminary floor clearing, and neither has a significant edge in cleaning area or speed. For those with limited budgets prioritizing cleaning effectiveness, other models like the T7s Plus might be a viable, more affordable alternative.

**type_one**: Home appliances

**type_two**: Home devices

**cnt_details**:
{ quality_score:2, share_count: 3, like_count: 8,, collect_count: 10, comment_count: 9 }

**user_question**: What are the differences between my original Xiaomi robot vacuum and the Roborock G10S in terms of usage?

**answer**: The Xiaomi robot primarily focuses on sweeping and has less advanced obstacle detection, often running over small objects. In contrast, the Roborock G10S offers better obstacle avoidance, customizable cleaning through its app, and combines sweeping with more effective mopping, though it initially struggles with accurate mapping and leaves more water residue.

**title**: Upgrading my new PS5 with a 2TB Fantom S790 SSD due to the price drop!

**contents**:
I recently got a PS5 Digital Edition and quickly realized the storage was a bit tight after installing a few games. I took the opportunity to sell my idle PS4 to get some cash back. Meanwhile, with SSD prices plummeting, I decided to go all out and bought the 2TB Fantom S790 solid-state drive. The drive comes in traditional Chinese black packaging with a highly recognizable Peking opera mask design. The back of the package lists the specific parameters and product information, offering a five-year warranty. Inside the box, there's the SSD itself, a small screwdriver, a screw, and a user manual. The drive has a graphene cooling sticker. If the heat sink uses a thermal silicone sheet, you'll need to remove the sheet before installation. The back of the drive also has a sticker with basic parameter info for easy reference.
Installation steps:

1. Before opening your console, lay something underneath to avoid scratches, then gently open the side of the case without the logo.

2. Use a Phillips screwdriver to open the SSD compartment at the bottom, which shows different mounting positions depending on the motherboard.

3. The screw and mount are fixed at the 110 position. Loosen the screw there.

4. This SSD is of the 2280 standard size, so set the mount at the 80 position.

5. Install the SSD, tighten the screw, reset the console to test it, then put the case back on.

Testing:

1. Power on, and the system directly detects the SSD and prompts for formatting.

2. I'm not sure if the speed is an issue.

3. After formatting, go to settings to select the default installation location for games.

4. You can choose the default installation location.

5. It shows a total of 2TB capacity, which is accurate


**type_one**: Computer and Technology

**type_two**:Computer Accessories

**cnt_details**:
{ quality_score:3, share_count: 4, like_count: 5,, collect_count: 27, comment_count: 15 }

**user_question**: Can I expand the storage on my Sony PlayStation 5 with a Fantom S790 SSD?

**answer**: Yes, you can use the Fanxiang S790 solid-state drive to expand the storage space of the Sony PlayStation 5 console. It should be noted that you need to follow the official tutorial and make sure that the solid-state drive used meets the requirements and standard specifications of the PS5.

## A.2 QA Generation with the help of GPT-4 API

In this study, we employ a methodology based on an assessment of a product review article to simulate user queries from the perspective of GPT-4. Utilizing the in-context learning approach (Ram et al., 2023), GPT-4 generates questions that mimic potential user concerns and queries. Subsequently, we apply a similar technique to enable the model to respond to these generated questions using information derived from the original article.

### A.2.1 Question Generation Prompt

Assume you are a customer purchasing goods on an e-commerce platform, and you have a question about a product. After reading the article, your question is well answered. What is your question?

Requirements: The question should be concise and end with a question mark "?"

[Example 1]
**Article**:
Practical test of the Xiaomi 65w gallium nitride charger charging a Samsung Note10+: Update modification (2020-03-06 17:53:12): Conclusion, the Xiaomi 65w gallium nitride does not support the 45w charging of Samsung Note10+. Among the several PPS-supporting PD chargers tested, the highest can reach the same 25w as the original charger.Xiaomi 65w Gallium NitrideWithout further ado, let's directly look at the pictures for yourself; the phone was in a screen-off state with about 35% battery level, using the original charger cable. For a few that did not have the original cable, I used the Lenovo lipstick cable.Baseus 65w Gallium Nitride
Green Union 30w PD
Anker 18w nano PD
Lenovo lipstick 65w PD
Set as main image
Xiaomi 65w PD (actually the same as the ZMI black 65w)
Samsung Note10+ original 25wPD
**Your question is**:
Can I use the Xiaomi 65w gallium nitride charger to charge my Samsung Note10+?

[Example 2]
**Article**:
{passage}
Your question is:

### A.2.2  Answer Generation Prompt

As a question-and-answer AI specializing in the e-commerce sector, I am equipped to answer user inquiries strictly based on the reference material provided. There are no scenarios in which I am unable to provide an answer, as all questions directly originate from the supplied references. Please rely on the information provided for accurate responses.

[Example 1]
**Article**:
Practical test of the Xiaomi 65w gallium nitride charger charging a Samsung Note10+: Update modification (2020-03-06 17:53:12): Conclusion, the Xiaomi 65w gallium nitride does not support the 45w charging of Samsung Note10+. Among the several PPS-supporting PD chargers tested, the highest can reach the same 25w as the original charger.Xiaomi 65w Gallium NitrideWithout further ado, let's directly look at the pictures for yourself; the phone was in a screen-off state with about 35% battery level, using the original charger cable. For a few that did not have the original cable, I used the Lenovo lipstick cable. Baseus 65w Gallium Nitride
Green Union 30w PD
Anker 18w nano PD
Lenovo lipstick 65w PD
Set as main image
Xiaomi 65w PD (actually the same as the ZMI black 65w)Samsung Note10+ original 25wPD.
**question** :
Can I use the Xiaomi 65w gallium nitride charger to charge my Samsung Note10+?
**Answer:**
You can use the Xiaomi 65w Gallium Nitride charger to charge your Samsung Note10+ smartphone; however, it is important to note that this charger does not support the 45w fast charging capability of the Samsung Note10+. According to test results, when utilizing the Xiaomi 65w Gallium Nitride charger, the maximum charging power attained is 25w, which is equivalent to the charging performance of the original Samsung Note10+ 25w PD charger. Therefore, although it cannot achieve the maximum charging speed of 45w, it can still safely charge at a rate of 25w.

[Example 2]
**Article**:
{passage}
**question**:
{question}
**Answer:**

Figure 5: Partial categories of WorthBuying dataset

## B   Model Training Details

### B.1   Domain-specific Continual Pretraining

In this section, we describe the continual pretraining process for the Baichuan2-7B-Base model (Yang et al., 2023) and the recipe of training data. Our models are trained on Llama-Factory (Zheng et al., 2024) framework.

#### B.1.1   Training Dataset

The detailed training datasets are listed in Table 7. For the e-commerce domain dataset, besides our WorthBuying dataset we collect e-commerce relevant notes from Xiaohongshu and Zhihu websites. For the generation domain data, we use the open-source pertaining dataset from Linly-OpenLLaMA [4] and WanJuan Corpus (Qiu et al., 2024)

#### B.1.2   Training Configurations

The model is trained using the DeepSpeed framework, which facilitates efficient distributed training. Key hyperparameters and configurations include:

- **Precision:** Training was performed using bfloat16 mixed precision, which significantly reduces memory consumption while maintaining the model's performance.

---

[4]https://github.com/CVI-SZU/Linly

| DataType | Language/Source | Weight | Number of Files | Size(GB) |
|---|---|---|---|---|
| | WorthyBuying Dataset | 2.53% | 735,975 | 2.43GB |
| E-commerce Domain dataset | Xhs-ecommerce-note | 4.96% | 2,841,686 | 4.5GB |
| | Zhihu-ecommerce-note | 7.91% | 2,996,802 | 7.59GB |
| | ZH-Wiki | 2.60% | 1,348,766 | 2.49GB |
| | Common-crawl-web | 20.29% | 6,741,652 | 19.44GB |
| General dataset | Chinese-new | 16.14% | 5,208,071 | 15.46GB |
| | Chinese-Law | 28.21% | 5,363,805 | 27.02GB |
| | Chinese-Patents | 15.79% | 1,127,932 | 15.13GB |
| | Chinese-CLS | 1.78% | 2,310,165 | 1.71GB |

Table 7: Training data of Continual pretraining

- **Optimizer:** We used the AdamW optimizer with a dynamically adjusted learning rate based on the training progression. This setup helps in stabilizing the training process in the initial phases and gradually fine-tuning the weights in later stages.

- **Learning Rate Scheduler:** A cosine decay scheduler with a warmup phase was implemented. The learning rate started at $2 \times 10^{-5}$ and was adjusted according to a predefined schedule to optimize convergence.

- **Model Parallelism:** We employed Zero Stage 2 optimization to manage GPU memory efficiently, offloading optimizer states and model parameters to CPU memory.

- **Hardware Configureation** The training was conducted on a cluster of four NVIDIA H100 80GB GPUs,

- **Other Configurations** We set 32 per device batch-size and use 100 step for warming up.

## B.2 Contrastive Learning Training Details

In our experiments, we utilize DeepSpeed for distributed training with four Nvidia-H100 80GB GPUs. Key hyperparameters included mixed precision training with bf16, a batch size of 4 per device with 4 gradient accumulation steps, and gradient checkpointing. The model is trained for one epoch with a learning rate of 1e-5, logging every 5 steps, and overwriting the output directory. Dataset processing uses 32 parallel processes, with cross-device negative sampling and a 100-step warmup phase to stabilize training.

In the training dataset, we integrate some open-source retrieval datasets with our WorthBuying dataset. For the open-source dataset, we use the train dataset of the retrieval part of C-MTEB (Xiao et al., 2023) Benchmark.

## C   Generation Evaluation Details

In the evaluation of generation, as mentioned in (Es et al., 2023), due to the subjective nature of responses generated by models, it is challenging to directly compute accuracy. Therefore, we adopt the multi-turn evaluation method utilized in *ragas* leveraging GPT-4. Below, we outline the prompts used in our evaluation to provide clarity and transparency in our assessment process.

## C.1 Stage1: Extracting key points from answer

Below are questions and answers, from which key points have been extracted and presented as individual statements.
========================

Question: What are the specifications of the Huawei X 14 smartphone?

Answer: Storage options include 256G and 512G; screen resolution is 2640x1080 FHD+ with a peak brightness of 1750 nits and a refresh rate of 120Hz; the internal screen measures 6.7 inches with a square aspect and 6.6 inches with rounded corners, while the external screen measures 3.4 inches square; it utilizes the Snapdragon 8 Gen 2 mobile platform with an octa-core CPU; the weight is a steady 187g; battery capacity is 3700mAh.

Statements:
Storage options of 256G and 512G.
Screen resolution of 2640x1080, refresh rate of 120Hz
Internal screen size of 6.7 inches, external screen of 3.4 inches square
Uses Snapdragon 8 Gen 2 mobile platform, CPU is octa-core
Weighs 187g.
Battery capacity is 3700mAh.
========================

Question: What oral supplements can help women restore their skin?

Answer: Vitamin C tablets from Northeast Pharmaceuticals, Kang'enbei's Vitamin C chewable tablets, and Northerland's compound Vitamin B and Vitamin C chewable tablets can help restore women's skin. Vitamin C can whiten the skin and regulate the body, while Vitamin B can repair the skin and is effective in treating acne and mouth ulcers.

Statements:

Vitamin C can whiten and help restore women's skin.
Vitamin B can repair the skin and is effective for acne and mouth ulcers.

========================

Question: Can the refill in a Baixue ballpoint pen be replaced?
Answer: Sorry, I cannot answer this question based on the available information.

Statement:

The model is unable to answer this question.

========================

Question: {question}
Answer: {answer}
Statement:

### C.2 Stage2: Judging the key points

Below are a series of statements along with corresponding facts. Each statement is evaluated for honesty, determining if it is sufficiently supported by the facts provided.
======================
Statement:
1. Storage capacity of 256 and 512GB.
2. Uses the Snapdragon 8 Gen 2 mobile platform, CPU has nine cores.
3. Weighs 199g.
4. Battery capacity is 3700mAh.

Fact:
This smartphone offers storage capacities of 256 and 512GB with an upgrade option available. Screen resolution is 2640x1080 FHD+ with a peak brightness of 1750 nits and a refresh rate of 120Hz; the internal screen sizes are 6.7 inches square and 6.6 inches rounded, external screen size is 3.4 inches square; it is equipped with the Snapdragon 8 Gen 2 mobile platform, and the CPU has eight cores; the weight is stable at 199g; battery capacity is 4600mAh.

Analysis:
1. Storage capacity of 256 and 512GB. This statement matches the facts, correct. [Yes]
2. Uses the Snapdragon 8 Gen 2 mobile platform, CPU has nine cores. According to the facts, the CPU should have eight cores, this statement is factually incorrect. [No]
3. Weighs 199g. This is consistent with the facts, correct. [Yes]
4. Battery capacity is 3700mAh. This does not match the facts, which state the battery capacity should be 4600mAh, incorrect. [No]

======================

Statement:
1. The model does not have enough information and cannot answer the question, the model's response is unknown.

Fact:
Vitamin C tablets from Northeast Pharmaceuticals, Kang'enbei's Vitamin C chewable tablets, and Northerland's compound Vitamin B and Vitamin C chewable tablets can help restore women's skin. Vitamin C has a whitening effect and can regulate the body, while Vitamin B can repair the skin and is effective in treating acne and mouth ulcers.

**Analysis:**
Although the model states it does not know, this makes it impossible for me to judge the response based on the provided facts. [Unknown]

======================

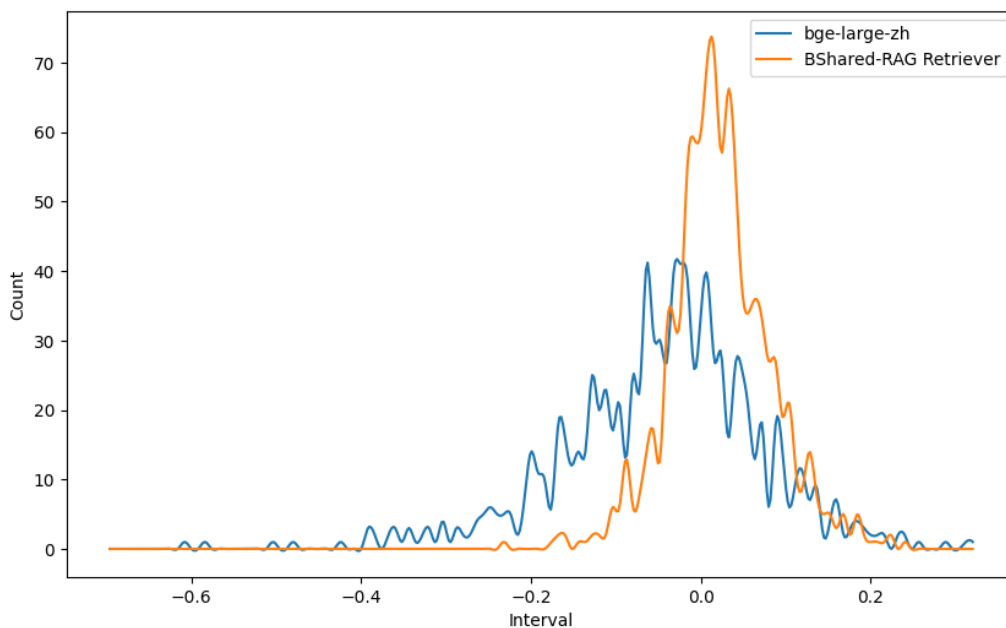Statement:
{statement}

Fact:
{ground_truth}

Analysis:

Figure 6: Histograms of gap scores, calculated by the similarity score of ground-truth document minus the similarity score of top returned not-ground-truth document, for two retrievers. The more positive gap scores the better.

## D    Additional Analysis

BShareRAG retriever can better distinguish hard negatives. We analyze the different behavior between two retrievers, *i.e.*, BGE-large-zh and BShared-RAG Retriever. For each question, we first calculate the gap of similarity scores between a ground-truth document (it surely contains answers) and the top returned not-ground-truth document. The gap score ranges from -1 to 1. A larger gap means a retriever can better rank the ground-truth document among others. Then we draw histograms of the gap scores for the two retrievers, as shown in Figure 6. It indicates that our BShared-RAG retriever can better distinguish the ground-truth document from other similar documents as the gap scores are positive for more questions. In contrast, *BGE-large-zh* has a more flat distribution, which means it cannot estimate a larger similarity of the ground-truth document for many questions.