processes

MDPI

Article

Domain-Specific Retrieval-Augmented Generation with Adaptive Embedding and Knowledge Distillation-Based Re-Ranking

Hao Luo [1,2,3], Xiong Luo [1,2,3,*], Weibo Zhao [1], Qiaojuan Peng [1,2,3], Ke Chen [1], Yinghui Liu [1] and Congcong Du [1]

[1] Beijing Key Laboratory of Demand Side Multi-Energy Carriers Optimization and Interaction Technique, China Electric Power Research Institute, Beijing 100192, China
[2] School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China
[3] Shunde Innovation School, University of Science and Technology Beijing, Foshan 528399, China
* Correspondence: xluo@ustb.edu.cn

Abstract

Retrieval-augmented generation (RAG) has emerged as an effective approach for analyzing massive and diverse data. It offers promising avenues for energy management and intelligent decision support amid the accelerating digital transformation of the power industry. However, when applied to this specialized domain, traditional RAG systems face two key challenges: (1) poor comprehension of domain-specific terminology, leading to irrelevant retrieval, and (2) limited precision in re-ranking the retrieved results. To address these limitations, this paper presents an innovative integrated optimization framework. The framework enhances RAG performance in the electric power domain through two key strategies. First, we adapt a base embedding model to the domain using contrastive learning and iteratively refine hard negative samples to improve retrieval quality. Second, we employ a large language model (LLM) as a teacher to distill re-ranking knowledge into a lightweight bidirectional encoder representations from transformers (BERT) model, using a hybrid loss function that combines mean squared error (MSE) loss and margin ranking loss. The framework aims to simultaneously improve the model's understanding of domain-specific terminology and the re-ranking accuracy of critical information. Experimental results on both a private power-domain dataset and the public DuReader_robust benchmark demonstrate that the proposed framework achieves significant performance gains. Comprehensive ablation studies confirm the necessity of each component and reveal their synergistic effects within the framework. Furthermore, sensitivity analyses of key hyperparameters confirm the effectiveness of our hybrid loss and identify optimal configurations that enhance both retrieval and generation performance. This work not only introduces an effective optimization framework tailored for domain-specific RAG applications but also advances industrial intelligence by enhancing the accuracy and reliability of information services.

Keywords: retrieval-augmented generation (RAG); fine-tuning; large language model (LLM); knowledge distillation; re-ranking optimization; contrastive learning

Academic Editor: Giancarlo Cravotto

Received: 1 December 2025
Revised: 18 December 2025
Accepted: 25 December 2025
Published: 27 December 2025

1. Introduction

With the continuous growth of global energy demand and the rapid adoption of renewable energy sources, the intelligent transformation of the power industry has become

an inevitable trend [1,2]. At its core lies the effective integration and in-depth analysis of massive, heterogeneous data to enable precise multi-dimensional load forecasting and to deliver scientifically grounded decision support. In this context, extracting actionable insights from complex data presents a significant challenge.

Recent advances in artificial intelligence—particularly breakthroughs in natural language processing (NLP) driven by large language models (LLMs) such as GPT-4 [3], LLaMA [4], and InstructGPT [5]—have opened new avenues for addressing critical information management challenges in the power sector, including the handling of vast volumes of technical documentation and the mitigation of information overload [6]. However, despite their impressive capabilities, LLMs are inherently constrained by their reliance on static training data and are prone to generating inaccurate or fabricated content—a phenomenon known as hallucination [7]. This limitation is especially problematic in knowledge-intensive and rapidly evolving domains like the power industry, where access to timely, accurate, and domain-specific knowledge is essential, and errors can lead to serious operational consequences [8].

To mitigate these issues and enhance both reliability and domain relevance, retrieval-augmented generation (RAG) has emerged as a promising paradigm [9]. By integrating external knowledge retrieval with text generation, RAG enables models to dynamically access up-to-date, authoritative information, thereby substantially reducing hallucinations and improving the accuracy and contextual appropriateness of generated responses.

However, applying RAG to specialized, knowledge-intensive domains presents significant challenges. General embedding models often fail to capture the intricate semantic nuances and specialized terminology inherent in domain-specific data (e.g., power industry documents), limiting retrieval results. Similarly, conventional ranking modules struggle to prioritize domain-critical information, compromising the precision and reliability of generated content. While prior works have optimized individual components, such as domain-adaptive embeddings [10–12] and ranking techniques [13,14], these efforts typically address retrieval or re-ranking in isolation. This piecemeal approach overlooks the interdependent nature of the RAG pipeline, where the quality of initial retrieval directly impacts the effectiveness of subsequent re-ranking. Consequently, a significant research gap exists in developing a holistic framework that co-optimizes both stages to maximize performance in specialized domains. Such a framework must first enhance the retriever's ability to surface semantically relevant candidates and then employ a re-ranker that refines these candidates with high precision, all while balancing performance with practical deployment needs.

To address this gap, we propose a domain-oriented RAG optimization framework that jointly adapts the retriever and the re-ranker for specialized corpora such as the power industry. The retriever is strengthened via contrastive learning with iterative hard-negative mining, and the re-ranker is trained by distilling an LLM teacher into a lightweight bidirectional encoder representations from transformers (BERT) model [15], using a novel hybrid loss function (mean squared error (MSE) loss and margin ranking loss).

Compared with prior multi-stage RAG optimization pipelines that (1) tune the retriever and the re-ranker as loosely coupled modules and (2) rely on task-specific heuristics to bridge stages, our framework enforces cross-stage training-signal consistency. Specifically, the retriever is adapted with a curriculum-style two-stage hard-negative mining (HNM) schedule that shapes the candidate distribution seen by the re-ranker, and the re-ranker is trained by distilling an LLM teacher with a hybrid objective that jointly preserves score calibration (regression to soft teacher scores) and relative ordering (pairwise margin constraints). This coupling targets a practical domain-RAG challenge: achieving

high-quality top-N re-ranking under limited domain annotations while keeping the final re-ranker deployable.

In summary, the contributions of this paper are as follows:

(1) We propose a novel integrated RAG framework specifically optimized for specialized knowledge domains, and demonstrate its efficacy within the power industry. This framework synergistically integrates domain-adaptive embedding fine-tuning with an innovative LLM-distilled, lightweight re-ranking mechanism, addressing the dual challenges of semantic comprehension and re-ranking precision in knowledge-intensive fields.

(2) We introduce a novel hybrid loss function (combining MSE loss and margin ranking loss) to facilitate the knowledge distillation from a powerful LLM to a lightweight BERT-based re-ranker. This distillation strategy allows the re-ranker to achieve superior ranking accuracy by inheriting the teacher's nuanced judgments. Crucially, the use of a lightweight student model provides a computationally practical alternative to deploying large, resource-intensive LLMs for re-ranking, making the approach more feasible for practical applications.

(3) We validate our framework through comprehensive experiments, demonstrating significant improvements in its retrieval and generation performance on different datasets. These improvements provide strong technical support for intelligent decision-making in the digital transformation of the power industry.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 details the proposed methodology; Section 4 describes the experimental setup, datasets, evaluation metrics, and presents the results analysis; and Section 5 concludes with a summary of our findings and discusses prospects for future work.

2. Related Work

RAG combines parameterized language models with external knowledge retrieval to enhance the accuracy and relevance of generated content, showing significant potential in knowledge-intensive domains [16]. Early RAG, often termed naive RAG, represents the most fundamental implementation, where a retrieval module fetches query-relevant documents from a knowledge base, and a generation module utilizes a pre-trained language model along with the retrieved context to produce answers [9]. While naive RAG has achieved initial success in open-domain question answering due to its simple structure and adaptability, it often struggles to capture the subtle semantics and specialized terminology inherent in domain-specific data, leading to unsatisfactory retrieval recall and generation quality.

To address these limitations, researchers have proposed a series of advanced RAG techniques that significantly enhance system performance by optimizing the retrieval process and subsequent context processing within the RAG pipeline. Regarding retrieval enhancement, for instance, the ITER-RETGEN method proposed by Shao et al. [17] iteratively alternates between generation and retrieval to progressively refine the acquired information. The FLARE method by Jiang et al. [18] proactively determines when and what to retrieve by predicting upcoming sentences to anticipate future information needs, enabling more dynamic and targeted retrieval. Similarly, the REPLUG method by Shi et al. [19] enhances black-box language models by integrating external retrieval, providing high-quality contextual augmentation without requiring model fine-tuning, making it particularly suitable for scenarios where internal model parameters are inaccessible. The Ret-Robust method by Yoran et al. [20] focuses on improving retriever robustness against noisy or adversarial queries. For context processing, Selective Context, proposed by Li et al. [21], trains a selector to identify the most critical information from a large pool of

retrieved passages, thereby enhancing generator efficiency and focus. Similarly, the SuRe method by Kim et al. [22] summarizes retrieved documents into concise answer candidates, improving the quality of context fed to LLMs for open-domain question answering. IRCoT, by Trivedi et al. [23], integrates chain-of-thought (CoT) prompting into an iterative retrieval framework to bolster evidence-based reasoning. Beyond optimizing initial retrieval and preliminary filtering, ensuring high-quality context for generation has made re-ranking a critical refinement stage in advanced RAG pipelines. Recent work leverages LLMs' semantic understanding to more accurately evaluate and select candidates [24,25].

However, retrieval and re-ranking are intrinsically interdependent: an advanced re-ranker cannot recover relevant evidence never recalled, while even a strong domain-adapted retriever may underperform due to suboptimal candidate ordering. This reveals a key gap—the lack of an integrated optimization strategy that jointly adapts retrieval and re-ranking to improve both domain-specific recall and ranking precision.

Recent re-ranking research has explored listwise encoding and contrastive objectives to better model inter-passage relationships. For instance, ListConRanker [26] introduces listwise encoding via a ListTransformer/ListAttention module and adopts Circle Loss for efficient training. In contrast, our re-ranker is lightweight and deployment-oriented: we distill an LLM teacher into a BERT-based cross-encoder with a hybrid objective combining score calibration and relative ordering signals. More importantly, our contribution extends beyond re-ranking: we couple retriever adaptation and re-ranker training through a curriculum-style hard-negative mining schedule to enforce cross-stage training-signal consistency.

3. Method

This section details the proposed methodology for optimizing a RAG specifically tailored for the complexities of the electric power domain. Standard RAG often struggles in specialized domains due to imprecise initial retrieval and suboptimal relevance ranking of retrieved candidates. Our approach systematically enhances the RAG pipeline by optimizing its core retrieval and re-ranking components, aiming to deliver more accurate and contextually relevant answers to domain-specific queries. The optimized workflow involves two primary enhancements: (1) the domain-adaptive fine-tuning of an embedding model through contrastive learning, which is further optimized by an iterative HNM strategy to improve initial retrieval; and (2) the training of a lightweight BERT-based re-ranker via knowledge distillation from an LLM, guided by a hybrid loss function combining MSE loss and margin ranking loss. These components are integrated into a RAG inference pipeline, detailed below.

3.1. Optimized RAG Framework

Our enhanced RAG framework, illustrated in Figure 1, modifies the standard RAG pipeline by integrating our two core contributions: a domain-adapted retriever and a distilled lightweight re-ranker. The inference process is as follows:

(1) Enhanced retrieval: Upon receiving a user query, it is first processed by our domain-adapted retriever. This retriever utilizes a text embedding model that has been fine-tuned specifically on electric power domain data (detailed in Section 3.2). This stage recalls an initial set of candidate passages from the knowledge corpus that are semantically relevant to the query, with improved accuracy over general-purpose embedding models.

(2) Distilled re-ranking: The retrieved candidate passages are then passed to a specialized re-ranker module. This module employs a lightweight BERT-based cross-encoder, trained via knowledge distillation (detailed in Section 3.3) to emulate the fine-grained relevance judgments of a powerful LLM. The re-ranker sorts these candidates, promoting the most pertinent passages to the forefront.

Finally, these top-N re-ranked, highly relevant passages are concatenated with the original query and fed as context to an LLM, typically along with a predefined prompt template, to synthesize the final answer.

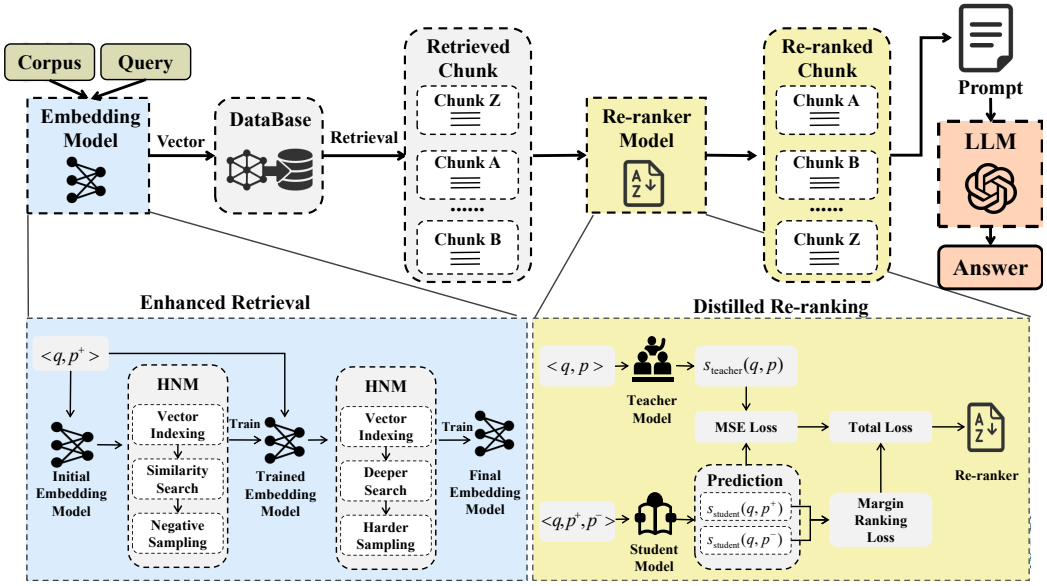

Figure 1. Overview of the optimized RAG framework with enhanced retrieval (blue, left) and distilled re-ranking (yellow, right) modules. In this diagram, q denotes a query, p^+ a positive (relevant) passage, and p^- a negative (irrelevant) passage. Chunks are labeled by their relative relevance rank (e.g., "A" denotes the highest-ranked passage after re-ranking) to illustrate the re-ordering process. Additionally, $s_{\text{teacher}}(q,p)$, $s_{\text{student}}(q,p^+)$, and $s_{\text{student}}(q,p^-)$ denote the normalized relevance probability scores of the teacher and student models for query-passage pairs.

3.2. Enhanced Retrieval: Domain-Adaptive Fine-Tuning

To improve the initial retrieval quality within the RAG, we fine-tune a base text embedding model for the electric power domain. This process employs contrastive learning, where the core idea is to learn an embedding space where queries are closer to their relevant (positive) passages and further from irrelevant (negative) passages. We use the information noise-contrastive estimation (InfoNCE) loss $\mathcal{L}_{\text{InfoNCE}}$:

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}\left[\log \frac{\exp(\text{sim}(q,p^+)/\tau)}{\exp(\text{sim}(q,p^+)/\tau) + \sum_{i=1}^{K}\exp(\text{sim}(q,p_i^-)/\tau)}\right], \quad (1)$$

where $\text{sim}(\cdot,\cdot)$ denotes the cosine similarity between the vector representations of the query q and the passage p, produced by the embedding model. p^+ is a positive passage relevant to q, p_i^- is the i-th negative passage, K is the total number of negative passages per query, and τ is a temperature hyperparameter.

The effectiveness of this contrastive learning heavily relies on the quality of negative samples [27]. To ensure the quality of these negatives, we employ a structured two-stage iterative HNM and fine-tuning process. As outlined in Algorithm 1, this approach is designed to progressively refine both the negative samples and the embedding model itself.

Algorithm 1 Two-Stage Retriever Fine-Tuning with Iterative HNM

Input: Document corpus $\mathcal{D}$; a set of labeled question-passage pairs $\mathcal{S} = \{(q_j, p_j^+)\}$; base text embedding model f; initial negative rank range $[L_1, U_1]$ (e.g., 5–20); optimized negative rank range $[L_2, U_2]$ (e.g., 3–15); number of candidates to retrieve for HNM, N.
Output: Domain-adapted fine-tuned embedding model f^*.

// Stage 1: Initial Domain Adaptation
Generate passage embeddings $\mathbf{V} = \{f(p) \mid p \in \mathcal{D}\}$.
Build search index $\mathcal{I}$ from $\mathbf{V}$.
Initialize training set $\mathcal{T}_1 = \emptyset$.
for each $(q_j, p_j^+) \in \mathcal{S}$ **do**
    Generate query embedding $\mathbf{q}_j = f(q_j)$.
    Retrieve top-N candidate passages $\mathcal{C}_j$ by searching $\mathcal{I}$ with $\mathbf{q}_j$.
    Construct initial hard negatives $\mathcal{P}_{j,1}^- = \{p_k \mid p_k \in \mathcal{C}_j, p_k \neq p_j^+,$ and $\text{rank}(p_k) \in [L_1, U_1]\}$.
    Add $(q_j, p_j^+, \mathcal{P}_{j,1}^-)$ to $\mathcal{T}_1$.
end for
Fine-tune f using training set $\mathcal{T}_1$ and $\mathcal{L}_{\text{InfoNCE}}$ from (1) to obtain f'.
// Stage 2: Optimized HNM and Further Fine-tuning
Generate new passage embeddings $\mathbf{V}' = \{f'(p) \mid p \in \mathcal{D}\}$.
Build new search index $\mathcal{I}'$ from $\mathbf{V}'$.
Initialize training set $\mathcal{T}_2 = \emptyset$.
for each $(q_j, p_j^+) \in \mathcal{S}$ **do**
    Generate new query embedding $\mathbf{q}_j = f'(q_j)$.
    Retrieve top-N candidate passages $\mathcal{C}_j$ by searching $\mathcal{I}'$ with $\mathbf{q}_j$.
    Construct more challenging hard negatives $\mathcal{P}_{j,2}^- = \{p_k \mid p_k \in \mathcal{C}_j, p_k \neq p_j^+,$ and $\text{rank}(p_k) \in [L_2, U_2]\}$.
    Add $(q_j, p_j^+, \mathcal{P}_{j,2}^-)$ to $\mathcal{T}_2$.
end for
Continue fine-tuning f' using training set $\mathcal{T}_2$ and $\mathcal{L}_{\text{InfoNCE}}$ to obtain f^*.
return f^*.

We adopt a two-stage hard-negative mining strategy to balance hardness and label noise. In early training, the retriever representation is still misaligned with the target domain; selecting extremely high-ranked negatives (e.g., ranks 1–3) tends to introduce false negatives that actually contain partial answers or near-duplicates, which destabilizes contrastive learning. Therefore, Stage 1 mines negatives from a slightly lower-ranked band (e.g., 5–20) that are semantically related yet less likely to be answer-bearing. After the model becomes domain-adapted, Stage 2 progressively shifts the mining band closer to the top (e.g., 3–15) to increase difficulty and improve discrimination among highly similar passages. This curriculum-like schedule improves training stability while still yielding strong hard negatives.

(1) Stage 1: Initial Domain Adaptation. This stage aims to provide the base text embedding model with foundational domain knowledge. We begin with the base text embedding model, denoted as f. First, we generate vector representations for all passages in the corpus $\mathcal{D}$ by encoding them with f, creating a vector set $\mathbf{V}$. A searchable index, $\mathcal{I}$, is then built from these vectors for efficient retrieval. For each question-passage pair (q_j, p_j^+) in our set of labeled pairs $\mathcal{S}$, we encode the query q_j and use its vector to search the index $\mathcal{I}$, retrieving the top-N candidate passages. From these candidates, we construct the initial hard negatives, $\mathcal{P}_{j,1}^-$, by selecting passages from a moderately distant rank range, $[L_1, U_1]$ (e.g., 5–20). This strategy provides a stable learning signal for the initial model, which still lacks domain expertise. These components form training triplets $(q_j, p_j^+, \mathcal{P}_{j,1}^-)$, which are collected

into a training set $\mathcal{T}_1$. Finally, the base text embedding model f is fine-tuned on $\mathcal{T}_1$ using the InfoNCE loss from (1), producing an intermediate model, f', tailored to the power domain.

(2) Stage 2: Optimized HNM. The second stage leverages the enhanced capabilities of the intermediate model f' to learn finer-grained distinctions. The process is repeated: we re-encode the entire corpus $\mathcal{D}$ using f' to generate a new, more accurate set of passage vectors and build a corresponding search index $\mathcal{I}'$. When retrieving candidates for each query q_j, we now mine for more challenging hard negatives, $\mathcal{P}^-_{j,2}$. These are selected from a rank range $[L_2, U_2]$ (e.g., 3–15) that is closer to the ground-truth positive passage p^+_j. Because f' already possesses domain awareness, these closer negatives are semantically very similar to the correct answer, compelling the model to learn more subtle and precise semantic differences. The model f' is then further fine-tuned using this new training set $\mathcal{T}_2$ of harder examples, which yields the final, optimized retriever model, f^*.

This iterative refinement, where the model itself helps identify progressively harder negative samples, is crucial for enhancing its discriminative capabilities. The resulting fine-tuned embedding model f^* serves as the optimized retriever in our RAG pipeline, responsible for efficiently recalling a high-quality set of initial candidate documents.

3.3. Distilled Re-Ranking: Learning from an LLM Teacher

To refine the top-N retrieved passages, we develop an efficient re-ranker via knowledge distillation from a fine-tuned LLM to a BERT-based cross-encoder. This ensures precise relevance ranking.

3.3.1. Teacher Model: Fine-Tuned LLM as a Pointwise Relevance Rater

We select a powerful LLM as the teacher, fine-tuning it as a pointwise relevance rater.

- Input and Output: The fundamental operation of the LLM teacher is to function as a scoring mechanism, evaluating the relevance of a text passage p to a given query q. During inference, the model processes their concatenated form as input. It then outputs a raw relevance logit, denoted as $S_{\text{teacher}}(q, p)$, which is subsequently normalized to produce a final relevance probability score, $s_{\text{teacher}}(q, p) \in [0, 1]$. We explicitly use the uppercase S for raw logits and the lowercase s for normalized probability scores to maintain clarity.
- Training Data: The LLM teacher is fine-tuned on triplets of (q, p, y), where q is a query, p is a passage, and $y \in \{0, 1\}$ is its ground-truth relevance label. Positive passages ($y = 1$) are directly derived from the relevant pairs in our labeled question-passage set $\mathcal{S}$. Crucially, negative passages ($y = 0$) are primarily sourced from the challenging hard negatives generated during the iterative retriever fine-tuning process (Section 3.2). This synergistic use of hard negatives ensures the teacher learns to discern subtle differences identified as difficult by the retriever.
- Fine-tuning Objective: The LLM is fine-tuned by minimizing the binary cross-entropy (BCE) loss $\mathcal{L}_{\text{BCE}}$:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{B}\sum_{i=1}^{B}[y_i \log(s_i) + (1 - y_i)\log(1 - s_i)], \tag{2}$$

where B is the batch size, y_i is the ground-truth binary label for the i-th training example, and s_i is the teacher LLM's predicted relevance probability for the i-th example.

This fine-tuned LLM teacher provides high-quality, "soft" relevance scores (probabilities $s_{\text{teacher}}(q, p)$) for subsequent knowledge distillation.

3.3.2. Student Model: BERT-Based Cross-Encoder

The student model is a lightweight BERT-based cross-encoder, designed for precise relevance re-ranking. It operates by taking a concatenated input sequence of the query and passage, typically formatted as "[CLS] q [SEP] p [SEP]". This sequence is processed by the BERT encoder to produce contextualized embeddings, from which a relevance logit $S_{\text{student}}(q, p)$ is extracted (e.g., from the [CLS] token's representation). This logit is then transformed into a relevance score $s_{\text{student}}(q, p)$ by applying a sigmoid function, $\sigma(S_{\text{student}}(q, p))$, where $\sigma(\cdot)$ denotes the sigmoid activation function.

3.3.3. Knowledge Distillation Training

The student re-ranker is trained to mimic the teacher using a hybrid loss on a distillation dataset of $(q, p, s_{\text{teacher}}(q, p))$ tuples, where $s_{\text{teacher}}(q, p)$ denotes the teacher's soft relevance score. The hybrid loss combines a regression-based objective and a relative ranking objective to address complementary aspects of re-ranking optimization. Specifically, the MSE term encourages score calibration by aligning the student's continuous outputs with the teacher's soft scores, providing smooth and globally consistent supervision. However, MSE alone does not explicitly enforce correct ordering among top-ranked candidates. The margin ranking loss directly optimizes relative ordering by enforcing a minimum margin between positive and negative passages, which is crucial for distinguishing highly similar candidates. By interpolating these two objectives with the weighting factor α, the hybrid loss balances global score calibration and local ranking discrimination, resulting in a student re-ranker that is both stable and ranking-consistent. The hybrid loss $\mathcal{L}_{\text{Hybrid}}$ is defined as:

$$\mathcal{L}_{\text{Hybrid}} = \alpha \cdot \mathcal{L}_{\text{MSE}} + (1 - \alpha) \cdot \mathcal{L}_{\text{MarginRanking}}. \tag{3}$$

where α is a hyperparameter to balance the contribution of the two losses.

- MSE loss ($\mathcal{L}_{\text{MSE}}$): This pointwise loss matches the student's predicted probability $s_{\text{student}}(q, p)$ to the teacher's probability $s_{\text{teacher}}(q, p)$ for any query-passage pair (q, p):

$$\mathcal{L}_{\text{MSE}} = \mathbb{E}_{(q,p)}\left[\left(s_{\text{teacher}}(q, p) - s_{\text{student}}(q, p)\right)^2\right]. \tag{4}$$

- Margin ranking loss ($\mathcal{L}_{\text{MarginRanking}}$): This pairwise loss learns relative ranking preferences. For a query q, a more relevant passage p^+, and a less relevant passage p^-, this loss encourages the student's score for p^+ to be greater than its score for p^- by at least a margin β (a positive hyperparameter):

$$\mathcal{L}_{\text{MarginRanking}} = \mathbb{E}_{(q,p^+,p^-)}\left[\max\left(0, \beta - \left(s_{\text{student}}(q, p^+) - s_{\text{student}}(q, p^-)\right)\right)\right]. \tag{5}$$

This hybrid approach ensures comprehensive knowledge transfer from the LLM teacher to the lightweight student model. The distilled BERT cross-encoder then serves as the second-stage re-ranker in our RAG pipeline, enabling precise re-ranking of candidates to enhance the overall quality of generated answers.

4. Experiment

This section conducts comprehensive experiments to evaluate the proposed method. It assesses the impact of our optimizations on the retrieval module's performance and the quality of generated answers, using both domain-specific and public datasets.

4.1. Datasets

4.1.1. Electric Power Question-Answering Dataset

To validate domain-specific performance, we construct a proprietary Chinese electric-power question-answering (QA) benchmark from internal industry documents covering data governance, power analysis and prediction, and business mechanism improvement. The dataset consists of a held-out test set for end-to-end evaluation and a separate training set for optimizing the retriever and the distillation-based re-ranker.

- Test Set: We develop 500 high-quality and standardized QA pairs by domain experts. Each instance includes: (1) a question, (2) a reference answer strictly grounded in the corpus, (3) a paragraph-level evidence span (reference information) that directly supports the answer, and (4) the source document identifier. This test set is used only for the final evaluation.
- Training Set: To train the domain-adaptive retriever and the knowledge-distillation re-ranker, domain experts construct 1000 training instances. Each instance contains one question, one positive passage, and 15 negative passages. In addition, we generate 1000 candidate training instances with an LLM and retain 685 instances after strict expert review and cleaning. The final training set therefore contains 1685 instances in total.
- Negative Sampling and Filtering: For each training question, negatives are selected as a mixture of (1) random/in-batch negatives, (2) topical negatives, and (3) mined hard negatives obtained via the iterative hard-negative mining procedure described in Algorithm 1. To reduce false negatives, we filter near-duplicate passages and remove candidate negatives that contain the answer string or overlap heavily with the annotated evidence span. Hard negatives follow the same two-stage mining schedule in Section 3.2 to progressively increase difficulty during training.
- Annotation Guidelines and Quality Control: We follow a standardized protocol: (1) each question must be answerable from a single paragraph-level evidence span in the corpus; (2) the reference answer must be strictly grounded in the evidence without external knowledge; (3) ambiguous or under-specified questions are discarded or rewritten; and (4) key entities and numerical values are normalized when applicable. Each instance is drafted by one annotator and reviewed by at least one senior domain expert; disagreements are resolved via adjudication. We additionally perform random spot checks and evidence–answer consistency validation to minimize labeling noise.
- Dataset Characterization and Bias Discussion: We report dataset statistics in Table 1 to assess potential annotation bias, including the distribution of question types (definition-based, procedural, and analytical) and the answer-length distribution (tokens). Potential biases may arise from expert curation (favoring canonical formulations and well-defined problems) and LLM-assisted generation (introducing templated expressions). We mitigate these effects via diversity constraints during construction, expert verification for generated samples, and coverage across multiple subdomains of power systems.
- Leakage Prevention: We split the proprietary dataset at the document level to avoid paragraph overlap between the training and test sets. LLM-assisted generation and screening are performed using training documents only, and we verify that test documents are not used during training data construction.

Table 1. Statistics of the proprietary power-domain QA dataset.

Item	Value
Training set	1685
Test set	500
Question types (Def./Proc./Anal.)	12.4% / 37.2% / 50.4%
Answer length (tokens, mean/median/p95)	137.1 / 119 / 291

Potential biases may arise from expert curation (favoring canonical formulations and well-defined problems) and LLM-assisted generation (introducing templated expressions or reducing linguistic diversity). Moreover, the relatively small dataset size may limit coverage of rare or highly ambiguous user intents. To mitigate these effects, we enforce diversity constraints during data construction, require expert verification for generated samples, and include questions spanning multiple subdomains of power systems. Despite these limitations, the dataset reflects realistic industrial QA use cases and provides a meaningful testbed for evaluating domain-adaptive retrieval and re-ranking methods.

We split the proprietary dataset at the document level to avoid paragraph overlap between training and test sets. LLM-assisted generation and screening are performed using training documents only, and we verify that test documents are not used during training data construction.

To complement the quantitative evaluation, we further analyze typical failure modes of the RAG pipeline.

Despite our rigorous quality control and leakage prevention measures during dataset construction, which ensure its robustness as an evaluation foundation, a deep understanding of the practical performance of the RAG (retrieval-augmented generation) pipeline is equally crucial. Therefore, to complement our quantitative evaluation and gain deeper insights into the specific challenges this pipeline encounters when interacting with our domain-specific data, we proceed with a qualitative error analysis. This analysis aims to uncover the typical failure modes that emerge when applying the RAG pipeline to our proprietary Chinese electric-power question-answering benchmark.

4.1.2. Error Analysis

We conduct a qualitative error analysis to better understand the failure modes of our RAG pipeline. We observe three representative patterns: (1) retrieval and re-ranking are correct, but the generator fails to use evidence (e.g., refusal or incomplete aggregation), (2) retrieval miss prevents downstream re-ranking and generation from recovering the correct answer, and (3) retrieval is correct, but re-ranking demotes the gold evidence below the context cutoff, leading to incorrect generation. Representative anonymized cases are provided in Appendix A.

4.1.3. Publicly Available Baseline Dataset

To evaluate the generalization ability of our proposed method, we also conduct experiments on the publicly available Chinese machine reading comprehension benchmark DuReader_robust [28]. The official dataset contains 15 K question-answer pairs for training, 1.4 K for validation, and 1.3 K for testing, with each item comprising a question, a reference document, and a short answer derived from Baidu search logs.

For our RAG experiments, we adapt the dataset as follows:

- Knowledge Corpus: We construct a unified knowledge base by merging all reference documents from the official training, validation, and test sets. This corpus serves as the external knowledge source for our RAG system to retrieve information.

- Training Set: We construct the training dataset by taking the 15 K questions from the official training set, using the passage containing the correct answer for each question as the positive sample, and generating 15 hard negatives per question from the DuReader_robust corpus via the HNM strategy described in Algorithm 1.
- Test Set: To evaluate the answer generation accuracy of the final RAG system, we use the 1.3 K question–answer pairs from the official test set.

Note that we do not use test questions/answers for training or tuning; the merged document set is treated as the fixed retrieval corpus, which is available to the system at inference time. This setup ensures that training and evaluation are separated at the level of supervised QA pairs (train vs. test questions/answers), while the merged document collection is treated as a fixed retrieval corpus available at inference time.

4.2. Assessment Indicators

To comprehensively evaluate the effectiveness of the proposed RAG framework, we adopt widely used evaluation metrics for both the main comparative experiments and the ablation studies. For generation quality, we employ ROUGE-1, ROUGE-2, ROUGE-L, and BLEU-4, which measure lexical overlap, fluency, and semantic relevance with respect to ground-truth responses. For retrieval performance, we use mean reciprocal rank (MRR) and normalized discounted cumulative gain (NDCG), which evaluate ranking accuracy and relevance ordering.

- ROUGE-1 [29]: Evaluates the recall of unigram overlaps between the candidate and reference answers, as shown in Equation (6):

$$\text{ROUGE-1} = \frac{\sum_{u \in C} \text{Count}_{\text{match}}(u, R)}{\sum_{u \in R} \text{Count}(u)}, \tag{6}$$

  where C is the candidate answer, R is the reference answer, $\text{Count}_{\text{match}}(u, R)$ is the number of occurrences of unigram u in C that also appears in R, and $\text{Count}(u)$ is the total number of occurrences of u in R.
- ROUGE-2: Measures the recall of bigram overlaps between the candidate and reference answers, reflecting phrase-level correspondence, as shown in Equation (7):

$$\text{ROUGE-2} = \frac{\sum_{b \in C} \text{Count}_{\text{match}}(b, R)}{\sum_{b \in R} \text{Count}(b)}, \tag{7}$$

  where b denotes a bigram.
- ROUGE-L: Quantifies sequence-level similarity by computing the longest common subsequence (LCS) between the candidate and reference answers, as defined in Equations (8)–(10):

$$R_{\text{LCS}} = \frac{\text{LCS}(C, R)}{|R|}, \tag{8}$$

$$P_{\text{LCS}} = \frac{\text{LCS}(C, R)}{|C|}, \tag{9}$$

$$\text{ROUGE-L} = \frac{(1 + \beta^2) R_{\text{LCS}} P_{\text{LCS}}}{R_{\text{LCS}} + \beta^2 P_{\text{LCS}}}, \tag{10}$$

  where $|C|$ and $|R|$ denote the lengths of C and R in tokens, and β balances precision and recall.

- MRR: Evaluates the average reciprocal rank of the first relevant document over all queries, as shown in Equation (11):

$$\text{MRR} = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{\text{rank}_q}, \tag{11}$$

where $|Q|$ is the number of queries, and rank_q denotes the position of the first relevant document for query q.

- BLEU-4 [30]: Computes the geometric mean of clipped n-gram precisions with uniform weights $w_n = 1/4$, adjusted by a brevity penalty, as shown in Equation (12):

$$\text{BLEU-4} = m \cdot \exp\left(\sum_{n=1}^{4} w_n \log p_n\right), \tag{12}$$

where p_n is the clipped n-gram precision, and the brevity penalty m is:

$$m = \begin{cases} 1, & \text{if } |C| > |R|, \\ \exp\left(1 - \frac{|R|}{|C|}\right), & \text{if } |C| \leq |R|. \end{cases} \tag{13}$$

- NDCG@k: Evaluates the quality of ranked retrieval results by considering both the relevance and position of documents up to rank k, as defined in Equations (14)–(16):

$$\text{DCG@}k = \sum_{i=1}^{k} \frac{2^{r_i} - 1}{\log_2(i+1)}, \tag{14}$$

$$\text{IDCG@}k = \sum_{i=1}^{k} \frac{2^{r_i^*} - 1}{\log_2(i+1)}, \tag{15}$$

$$\text{NDCG@}k = \frac{\text{DCG@}k}{\text{IDCG@}k}, \tag{16}$$

where r_i is the graded relevance score at rank i, and r_i^* is the graded relevance score at rank i in the ideal ranking.

4.3. Experimental Setup

Our baseline experiments are conducted on the FlashRAG [31] framework. The experiments are conducted on an NVIDIA A800 GPU with 80 GB of GPU memory. The generative LLM in the RAG is uniformly set to Llama3-Chinese-8B-Instruct (https://huggingface.co/hfl/llama-3-chinese-8b-instruct (accessed on 24 December 2025)), and the embedding retrieval model is bge-large-zh-v1.5 (https://huggingface.co/BAAI/bge-large-zh-v1.5 (accessed on 24 December 2025)). In the distillation experiments, we use Qwen2.5-3B-Instruct (https://huggingface.co/Qwen/Qwen2.5-3B-Instruct (accessed on 24 December 2025)) as the teacher re-ranker, and chinese-roberta-wwm-ext (https://huggingface.co/hfl/chinese-roberta-wwm-ext (accessed on 24 December 2025)) as the student re-ranker.

To enable a controlled comparison, we implement and run all methods within the same FlashRAG framework and keep the system components identical as much as possible except for the "retrieval–generation strategy", including the corpus, chunking strategy, embedding retrieval model (bge-large-zh-v1.5), indexing pipeline, the generative LLM, and decoding hyperparameters.

For experiments that involve re-ranking, we set the number of initially retrieved documents (retriever_top) to 10 and the number of documents kept after re-ranking (rerank_top) to 5, unless otherwise specified in dedicated sensitivity analyses where retriever_top is

intentionally varied to study its impact (see Section 4.6.3). For experiments that rely on basic embedding-based retrieval only (i.e., retriever-only without re-ranking), we set retriever_top to 5, which is also the final number of documents passed to the generator.

For methods that inherently require multi-step reasoning or chain-of-thought style prompting (e.g., IRCoT), the prompt template is considered part of the method definition. Therefore, we use Chinese-equivalent translations of the prompts from the original papers and keep them fixed across all datasets. Beyond necessary localization into Chinese, we do not perform additional manual prompt tuning for specific methods or datasets.

4.4. Comparison of Methods

Tables 2 and 3 show the comparison of results between our method and other methods on the power sector dataset and the DuReader_robust dataset, respectively. The results in Tables 2 and 3 show that our method achieves the best overall performance among the compared approaches under our unified experimental setting, improving all four metrics (ROUGE-1/2/L and BLEU-4) on both datasets. Specifically, when compared to the strongest performing counterpart, ITER-RETGEN, our method demonstrates improvements on the electric power dataset in ROUGE-1, ROUGE-2, ROUGE-L, and BLEU-4 by 2.61, 2.91, 2.47, and 3.00, respectively. On the DuReader_robust dataset, the corresponding enhancements are even more substantial, at 16.47, 12.72, 16.55, and 8.25. The performance gains relative to the baseline experiment are markedly more significant.

Table 2. Experimental results on the electric power dataset.

Method	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-4
Naive [1]	53.67	37.30	44.94	25.08
ITER-RETGEN [17]	54.31	38.05	45.94	25.76
FLARE [18]	45.73	29.93	37.46	17.76
REPLUG [19]	40.49	20.61	31.58	10.52
Ret-Robust [20]	47.94	29.74	38.00	19.47
Selective Context [21]	43.26	22.16	33.56	11.61
SuRe [22]	48.45	30.09	39.89	19.39
IRCoT [23]	50.88	33.45	42.58	22.39
ListConRanker [26]	47.73	29.86	38.75	19.28
Our method	56.92	40.96	48.41	28.76

[1] Naive denotes the baseline RAG system that employs the bge-large-zh-v1.5 embedding model without domain-specific fine-tuning and a re-ranking component, serving as a reference for evaluating optimization strategies.

Table 3. Experimental results on the DuReader_robust dataset.

Method	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-4
Naive [1]	40.93	24.71	40.62	18.03
ITER-RETGEN [17]	41.68	25.48	41.35	18.10
FLARE [18]	39.26	22.72	38.11	14.24
REPLUG [19]	24.26	9.60	24.04	7.34
Ret-Robust [20]	41.03	23.00	40.87	17.05
Selective Context [21]	33.35	17.30	32.90	11.80
SuRe [22]	38.97	23.92	38.63	17.85
IRCoT [23]	38.71	24.31	38.44	16.95
ListConRanker [26]	42.70	29.04	42.25	18.28
Our method	58.15	38.20	57.90	26.35

[1] Naive denotes the baseline RAG system that employs the bge-large-zh-v1.5 embedding model without domain-specific fine-tuning and a re-ranking component, serving as a reference for evaluating optimization strategies.

The more pronounced effect observed on the DuReader_robust dataset may be attributed to the fact that its answers are sufficiently direct and concise. In such scenarios, if a RAG fails to retrieve the most critical contextual passages, *n*-gram-based metrics such as ROUGE can experience a sharp decline. Conversely, the reference answers within the electric power dataset are generally more detailed and descriptive, encompassing a wealth of contextual information. Consequently, even if the documents retrieved are not optimally aligned, they may still contain overlapping terminology or semantically similar expressions to the standard answers. This can lead to ROUGE scores that are already relatively high, even for baseline systems, thereby making the numerical gains from subsequent optimizations appear less prominent, despite potential substantive improvements in the actual quality of information retrieval and ranking.

These findings robustly validate the efficacy of our proposed combined optimization framework within the target-specific domain, as well as its potential generalization capabilities. Through targeted optimization, our method exhibits an enhanced capacity to comprehend specialized knowledge within the electric power domain, select more pertinent context, and generate answers that are more accurate and align more closely with domain-specific characteristics. Furthermore, it proves equally effective in advancing the performance of RAG in general-purpose domains.

4.5. Ablation Experiments

To gain a deeper understanding of the contribution of each key technique within our framework, we conduct a series of ablation experiments on the electric power dataset. These experiments start from the naive configuration (as defined in the comparative experiments), which utilizes only a basic embedding model for retrieval without any re-ranking functionality. Against this baseline, we systematically test the impact of removing or replacing specific modules, including: using only the fine-tuned retriever (FT Ret.), using the base retriever with the distilled re-ranker (Base Ret. + Distilled RR), using the base retriever with a basic, non-distilled re-ranker (Base Ret. + Basic RR), the full method without fine-tuning the LLM (w/o LLM FT), and the full method without iterative HNM (w/o Iterative HNM). We evaluate the impact of these changes on end-to-end generation quality using ROUGE-1, ROUGE-2, ROUGE-L, and BLEU-4, and on retrieval and ranking performance using MRR and NDCG@5. The results are presented in Table 4.

Table 4. Ablation experiments on the electric power dataset.

Method	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-4	MRR	NDCG@5
Naive	53.67	37.30	44.94	25.08	39.71	41.51
FT Ret.	54.92	38.89	46.31	26.86	45.95	47.63
Base Ret. + Distilled RR	55.59	38.69	46.76	26.81	57.57	56.28
Base Ret. + Basic RR	48.62	30.62	39.11	19.64	17.72	20.74
w/o LLM FT	55.08	38.95	46.47	27.81	46.82	49.98
w/o Iterative HNM	55.46	39.40	47.06	27.47	47.89	50.84
Our method	56.92	40.96	48.41	28.76	64.53	63.92

- Domain-adapted retriever boosts performance: Fine-tuning the retriever alone (FT Ret.) demonstrates significant gains over naive across all metrics, validating the benefit of domain adaptation for the retriever.
- Distilled re-ranker shows strong independent impact: Our knowledge-distilled re-ranker, even with a base retriever (Base Ret. + Distilled RR), substantially improves upon naive (ROUGE-L: 46.76, MRR: 57.57), demonstrating its effectiveness in refining relevance judgments and achieving strong ranking scores (NDCG@5: 56.28).

- Knowledge distillation is crucial for re-ranking: A basic, non-distilled re-ranker (Base Ret. + Basic RR), implemented using the chinese-roberta-wwm-ext model in our experiments, severely degrades performance (ROUGE-L: 39.11, MRR: 17.72), falling far below naive (ROUGE-L: 44.94, MRR: 39.71). This degradation is attributable to the model's limited parameter size and architectural constraints, which impair its reasoning and relevance judgment capabilities during re-ranking. These results highlight the importance of knowledge distillation in enhancing the effectiveness of lightweight re-rankers, as a non-distilled variant underperforms in this specific setup.
- Quality of the LLM teacher is paramount: Removing the LLM fine-tuning (w/o LLM FT) results in a moderate performance decline (ROUGE-L: 46.47, MRR: 46.82), with results above naive but below the full method, underscoring the necessity of a high-quality, domain-aware teacher for successful distillation.
- Iterative HNM contributes measurably: Disabling iterative HNM (w/o Iterative HNM) results in a noticeable performance drop compared to the full method (ROUGE-L: 47.06, MRR: 47.89), confirming the value of this iterative refinement strategy for retriever robustness.
- Full method achieves synergistic superiority: Our complete proposed method (our method), integrating all optimizations, consistently achieves the best results across all metrics. This highlights a clear synergistic effect and validates the overall design of our multi-stage framework.

4.6. Sensitivity Analysis Experiment

In this subsection, we conduct a series of sensitivity analyses to investigate the impact of key hyperparameters within our proposed framework. All experiments presented here are performed on the electric power dataset.

4.6.1. The Impact of Hybrid Loss Weighting in Knowledge Distillation

The hybrid loss function for the re-ranker combines MSE and margin ranking loss, with the weight α controlling their balance (see Equation (3)). Table 5 shows the performance impact of varying α on the electric power dataset. The results reveal a clear trend: starting from a pure margin ranking loss ($\alpha = 0$), the generation scores for all ROUGE and BLEU-4 metrics begin to rise as the MSE loss is introduced. This improvement continues until performance peaks at $\alpha = 0.5$, where all four metrics achieve their best scores (ROUGE-1 at 56.87, ROUGE-2 at 40.64, ROUGE-L at 48.08, and BLEU-4 at 28.36). Beyond this optimal point, as the α value continues to increase towards 1.0, the delicate balance between the two complementary loss components is disrupted. This over-reliance on a single aspect of the training objective causes the scores to consistently decline. This demonstrates that a balanced hybrid approach significantly outperforms a model trained on a single loss function for downstream generation tasks.

The effect of α on the retrieval metrics is visualized in Figure 2. Both the MRR and NDCG@5 metrics exhibit a clear and similar trend. Figure 2 shows that at $\alpha = 0$ (representing pure margin ranking loss), performance is already strong. As a small amount of the MSE distillation loss is introduced (i.e., $\alpha > 0$), both metrics begin to climb, indicating a positive contribution from the distillation objective. As illustrated by the plot, both curves reach their apex in the middle of the range, achieving their optimal performance around $\alpha = 0.5$. Beyond this point, as α increases towards 1.0, both MRR and NDCG@5 show a steady decline.

Table 5. Sensitivity analysis for the distillation loss weight (α) in the re-ranker's hybrid loss function.

α	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-4
0	55.97	39.93	47.55	27.89
0.1	56.60	40.37	47.99	28.08
0.2	56.36	40.12	47.70	28.05
0.3	56.46	39.99	47.84	28.00
0.4	56.36	40.16	47.83	28.18
0.5	56.87	40.64	48.08	28.36
0.6	56.80	40.52	48.01	28.24
0.7	56.66	40.23	47.88	27.85
0.8	56.60	40.23	47.97	28.07
0.9	55.97	39.74	47.40	27.70
1.0	56.23	40.04	47.59	27.96

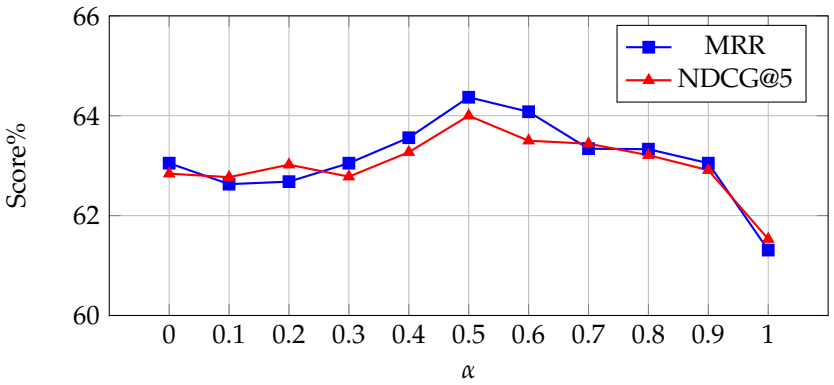

Figure 2. Sensitivity analysis of the re-ranker's distillation loss weight (α) on retrieval metrics.

Taken together, these results confirm that an α value of 0.5 strikes the optimal balance for both generation and retrieval tasks. This validates our hypothesis that the two loss functions provide complementary benefits. The pointwise loss function ($\mathcal{L}_{\text{MSE}}$) encourages the student model to learn the absolute strength of the teacher's judgment on each document's relevance, which helps the model achieve better calibration and understand subtle differences. Meanwhile, the margin ranking loss ($\mathcal{L}_{\text{MarginRanking}}$) directly optimizes the model's relative ranking ability, ensuring that more relevant documents are ranked higher. The hybrid loss function successfully combines the advantages of both methods, enabling the student re-ranker to more comprehensively and effectively absorb the teacher model's knowledge, ultimately achieving high-performance re-ranking.

4.6.2. The Impact of Margin in Margin Ranking Loss

We then conduct a sensitivity analysis on the margin hyperparameter β in the Margin Ranking Loss component of our hybrid loss function. To isolate the impact of the margin, we hold the distillation loss weight fixed at $\alpha = 0.3$ for this set of experiments. We evaluate the effect of β on both generation quality, measured by the ROUGE suite (ROUGE-1, ROUGE-2, and ROUGE-L) and BLEU-4, and on retrieval effectiveness using MRR and NDCG@5. The value of β is varied from 0.1 to 1.0, with the results for generation metrics shown in Table 6 and for retrieval metrics in Figure 3.

Table 6. Sensitivity analysis of the β in margin ranking loss on generation metrics.

β	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-4
0.1	56.28	40.01	47.78	28.10
0.2	56.65	40.77	48.09	28.62
0.3	56.46	39.99	47.84	28.00
0.4	56.92	40.96	48.41	28.76
0.5	56.39	40.23	47.75	28.09
0.6	56.35	40.11	47.88	27.96
0.7	56.37	40.18	47.82	28.22
0.8	56.22	40.02	47.57	27.92
0.9	56.02	39.52	47.31	27.39
1.0	55.38	38.82	46.65	26.82

Table 6 reveals the influence of the margin β on generation quality. A distinct "increase-then-decrease" pattern is evident. Starting from a small margin of $\beta = 0.1$, all generation metrics climb to their peak performance at $\beta = 0.4$ (ROUGE-1 at 56.92, ROUGE-2 at 40.96, ROUGE-L at 48.41, and BLEU-4 at 28.76). As the margin is increased further beyond this optimal point, performance consistently declines, with a particularly sharp drop at $\beta = 1.0$. This indicates that a moderate margin of 0.4 is most effective for training the re-ranker to produce context that maximizes downstream generation quality.

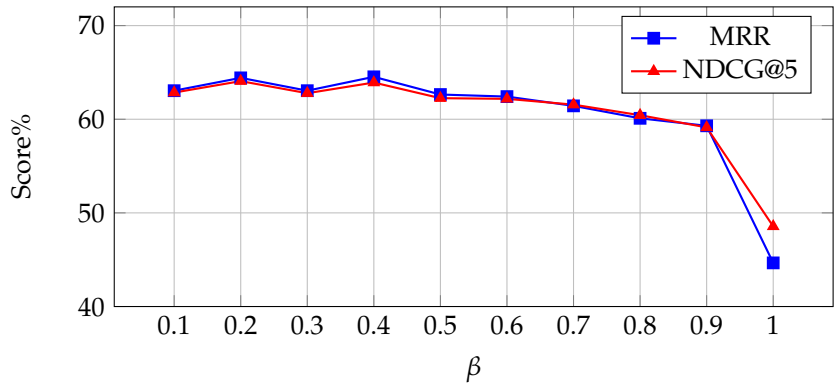

Figure 3. Impact of the β in margin ranking loss on retrieval performance.

Figure 3 illustrates the effect of the β on retrieval performance. Both metrics show strong performance in the lower β range. MRR achieves its highest score at $\beta = 0.4$, with $\beta = 0.2$ being a close second. NDCG@5 peaks earlier at $\beta = 0.2$, with $\beta = 0.4$ being the second-best value. Both metrics exhibit a gradual decline for β above 0.4, followed by a sharp decrease when β reaches 1.0. This suggests that β between 0.2 and 0.4 is optimal for maximizing these traditional retrieval metrics, effectively training the model to re-rank relevant documents highly.

4.6.3. The Impact of the Number of Retrieved Documents

To investigate the interplay between the retriever and re-ranker, we analyze the system's sensitivity to the number of initially retrieved documents, retriever_top. For this set of experiments, the re-ranker's hyperparameters are held fixed with the distillation loss weight set to $\alpha = 0.3$ and the margin ranking loss margin to $\beta = 0.4$. We keep the final number of documents sent to the generator fixed at rerank_top = 5 and evaluate the system with retriever_top values of 5, 10, 15, 20, 25, and 30. The results, encompassing both generation and retrieval metrics, are presented in Table 7.

Table 7. Sensitivity analysis of the number of initially retrieved documents (retriever_top), with rerank_top fixed at 5.

Retriever_Top	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-4	MRR	NDCG@5
5	55.13	39.01	46.60	27.15	56.59	56.48
10	56.92	40.96	48.41	28.76	64.53	63.92
15	58.05	42.05	49.18	29.35	64.37	64.00
20	57.38	41.19	48.80	29.01	61.41	62.20
25	57.18	41.06	48.44	28.68	61.83	62.18
30	57.33	41.27	48.81	28.93	62.34	62.74

The experimental results in Table 7 reveal a nuanced relationship between the size of the initial candidate pool and the final system performance. We observe a clear "increase-then-decrease" trend across both generation and retrieval metrics, with performance peaking at retriever_top = 15 for generation metrics (ROUGE and BLEU-4) and at retriever_top = 10 or 15 for retrieval metrics (MRR and NDCG@5).

Specifically, the end-to-end generation quality, measured by ROUGE and BLEU, reaches its undisputed peak at retriever_top = 15. This indicates that a candidate pool of 15 documents strikes the optimal balance, providing a rich enough source for our re-ranker to construct a context set with high informational synergy and diversity, thereby empowering the generator to produce the most accurate and comprehensive answers. When the pool size is smaller (retriever_top = 5), critical information may be missed during the initial retrieval phase. Conversely, when the pool becomes too large (retriever_top = 25), the increase in redundant or noisy documents causes "context noise", which impairs the quality of the final context set and thus reduces the accuracy of the generated output. The retrieval metrics (MRR and NDCG@5) achieve their highest values at retriever_top = 10 or 15, followed by a modest decline at higher values. This demonstrates a realistic limitation of any re-ranking model: as the candidate pool grows larger and noisier, the re-ranker's task of precisely identifying the single most relevant document becomes more challenging due to the increased presence of distracting, highly similar negative samples.

These findings underscore the importance of balancing the initial retrieval pool size in domain-specific RAG frameworks, such as ours, where the domain-adapted retriever (fine-tuned via contrastive learning and iterative hard negative mining) and the distilled re-ranker (optimized with a hybrid loss function) work synergistically to enhance overall performance. Optimal configurations like retriever_top = 15 not only improve generation quality but also provide practical guidance for deploying efficient RAG systems in the electric power industry.

5. Conclusions

This study focuses on addressing the challenges faced by existing RAG systems when dealing with knowledge-intensive and specialized domains, exemplified by the electric power industry. To enhance the accuracy of information retrieval and the reliability of generated answers, thereby better serving the needs of the electric power industry, this study proposes and validates a multi-stage RAG optimization framework. The framework systematically integrates innovations in the following key technologies:

(1) Domain-adaptive embedding representation learning: Utilizing an iterative HNM strategy combined with contrastive learning to fine-tune the embedding model. This process significantly improves the model's capacity to understand domain-specific terminology and differentiate between complex semantics within the electric power industry, leading to more precise document retrieval.

(2) Knowledge distillation for re-ranking: Building upon the enhanced retriever, we introduce a novel re-ranking mechanism. We utilize a powerful LLM as a teacher and distill its sophisticated ranking knowledge into a lightweight BERT-based student model. This process is guided by a carefully designed hybrid loss function that combines MSE loss and margin ranking loss, providing a computationally practical yet highly effective re-ranking solution.

Through extensive experiments on a private electric power domain dataset and the publicly available DuReader_robust dataset, our proposed method demonstrates significant improvements in end-to-end generation compared to existing approaches, along with strong generalization capability. Exhaustive ablation experiments and hyperparameter sensitivity analyses further confirm the necessity and effectiveness of each component within the framework.

For future research, we plan to explore several promising directions. First, we will investigate adaptive fine-tuning strategies for the teacher LLM to reduce its sensitivity to model selection and enhance distillation efficiency. Second, to address the challenge of dynamic knowledge in real-world scenarios, we aim to integrate real-time data streaming and incremental learning techniques, which would enable the framework to handle evolving knowledge bases without complete retraining. Finally, we will research methods for improving the framework's zero-shot or few-shot generalization capabilities, allowing for effective deployment in new specialized domains with minimal adaptation. These efforts will further strengthen the framework's applicability to real-world, knowledge-intensive applications.

Author Contributions: Conceptualization, X.L. and H.L.; methodology, H.L. and Q.P.; software, H.L. with assistance from W.Z.; validation, C.D., K.C. and Q.P.; formal analysis, H.L. and Q.P.; investigation, H.L., C.D. and W.Z.; resources, Y.L., K.C. and W.Z.; data curation, H.L. and Y.L.; writing—original draft preparation, H.L.; writing—review and editing, Q.P., X.L. and H.L.; visualization, H.L. with support from C.D.; supervision, X.L. and Q.P.; project administration, Y.L., K.C. and X.L.; funding acquisition, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work is financially supported by the 2025 Open Fund of Beijing Key Laboratory of Demand Side Multi-Energy Carriers Optimization and Interaction Technique (project title: Research on multi-source cognition-driven intelligent modeling and generative reasoning techniques for user load forecasting, Project No. SGDK0000YDJS2504840-B1).

Data Availability Statement: DuReader_robust is a publicly available benchmark and can be accessed from its official release. The proprietary Chinese power-domain QA dataset used in this study cannot be publicly released due to industrial confidentiality and data governance constraints.

Acknowledgments: All authors who contributed to this study are gratefully acknowledged.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Qualitative Failure Cases

This appendix presents representative failure cases of our method to complement the qualitative analysis in the main text. All examples are anonymized and simplified for clarity. We summarize retrieved evidence rather than reproducing full passages to avoid revealing sensitive content. All Chinese text in this appendix is accompanied by an English translation for clarity and accessibility to international readers.

C1 (Retriever✓, Re-ranker✓, Generator×).

Question: What are the main conclusions and recommendations of the study on Henan's key industrial chains and their adjustable capacity?

Gold answer (key points): The answer should cover four conclusions/recommendations regarding (1) industrial chain mapping and layout, (2) manufacturing sector's peak-shaving

capability during summer evenings and the "5% peak reduction" target, (3) a coordinated invitation-based demand response mechanism, and (4) the potential for air-conditioning load regulation.

Gold evidence (summary):

- The industrial chain map and development layout have been clarified, enabling enterprise-level analysis across the entire chain.
- Manufacturing sectors have limited adjustability during summer evening peaks; achieving the "5% peak reduction" requires multi-stakeholder efforts.
- Enterprises exhibit high economic interdependence; a coordinated invitation-based demand response mechanism is recommended.
- Residential and commercial cooling loads are substantial during evening peaks, offering significant potential for air-conditioning load control.

System evidence snapshot: The retriever and re-ranker include the gold evidence in the final context (top-5 passages).

Model output (incorrect): Based on the provided information, it is not possible to determine the main conclusions and recommendations of the study on Henan's key industrial chains and their adjustable capacity.

Diagnosis: Although multiple relevant passages were correctly retrieved and ranked within the top-N candidates, the generator failed to synthesize the correct answer. This failure is primarily attributed to semantic fragmentation during chunking. The critical information required for a complete answer was inadvertently split across different text chunks during the preprocessing phase. Consequently, while individual chunks were recalled, none of them contained the full semantic unit, leading to an incomplete understanding by the LLM. This, combined with the insufficient reasoning capability of the LLM to effectively integrate these fragmented semantic units into a coherent and complete answer, resulted in a conservative "refusal to answer" response.

C2 (Retriever×).

Question: What are the responsibilities of the State Grid Digitalization Department?

Gold answer (key points): Key responsibilities include: (1) supporting data sharing (e.g., external meteorological and economic data); (2) co-leading platform development and the construction of the "three repositories" with the Marketing Department; (3) conducting research on forecasting models; (4) centralized management of platform operation and maintenance; and (5) providing hardware and software infrastructure for forecasting.

Gold evidence (summary):

- Orchestrating external data sharing services to support electricity demand and generation forecasting;
- Co-leading with the Marketing Department the development of the platform and the "three repositories": model event repository, user sample repository, and experience/-knowledge repository;
- Collaboratively conducting research on electricity demand and generation forecasting models;
- Centralized oversight of platform operation and maintenance;
- Providing the necessary hardware and software infrastructure for forecasting.

System evidence snapshot: The gold evidence is not retrieved in top-5 candidates ($K = 5$); retrieved passages mainly describe general digital platform application and business digital transformation.

Model output (incorrect): The responsibilities of the State Grid Digitalization Department include enhancing business support capabilities of system platforms, promoting platform integration and data sharing, fully implementing mobile operational services, strengthening

data resource governance, and comprehensively supporting the digital transformation of market-oriented business operations.

Diagnosis: This represents a semantic retrieval failure. The embedding model failed to distinguish between the specific administrative "duties of the Digitalization Department" and general "tasks for deepening digital platform applications." Due to high semantic overlap in keywords (e.g., "digitalization," "platform support"), the retriever recalled documents describing general technical work content rather than specific organizational responsibilities, causing the generator to hallucinate a response based on irrelevant project descriptions.

C3 (Retriever✓, Re-ranker×).

Question: What is the objective of China's national carbon market?

Gold answer (key points): The answer should clearly state the core objective of China's national carbon market—reducing the carbon intensity of economic activities, i.e., lowering average emissions per unit output from covered facilities—and explain that this is achieved through compliance-period verification and allowance adjustment.

Gold evidence (summary):

- The national carbon market commenced trading in July 2021;
- Objective: to reduce the carbon intensity of economic activities (i.e., decrease average emissions per unit of output from covered entities);
- After each compliance period, actual output is verified and final emission allowances are adjusted accordingly.

System evidence snapshot: The fine-tuned retriever retrieves the gold evidence within the top-K candidates, but the re-ranker demotes it below the final context cutoff (e.g., moved from rank-2 to rank-9), while promoting passages about trading procedures/compliance details that do not explicitly state the goal.

Model output (incorrect): The goal of China's carbon market is to essentially establish a standard system for carbon peaking and carbon neutrality.

Diagnosis: This is a re-ranking alignment error. Although the retriever successfully recalled the correct document regarding the carbon market's economic goals, the re-ranker incorrectly demoted it. Instead, the re-ranker assigned higher scores to a distractor document concerning the "Guidelines for Carbon Peaking and Carbon Neutrality Standard System Construction." The re-ranker was misled by the strong semantic similarity of "carbon targets" in the fundamental principles of the standard system, failing to differentiate them from the specific operational goals of the carbon trading market.

References

1. Ieva, S.; Loconte, D.; Loseto, G.; Ruta, M.; Scioscia, F.; Marche, D.; Notarnicola, M. A Retrieval-Augmented Generation Approach for Data-Driven Energy Infrastructure Digital Twins. *Smart Cities* **2024**, *7*, 3095–3120. [CrossRef]
2. Pandey, U.; Pathak, A.; Kumar, A.; Mondal, S. Applications of artificial intelligence in power system operation, control and planning: A review. *Clean Energy* **2023**, *7*, 1199–1218. [CrossRef]
3. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774. [CrossRef]
4. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMa: Open and Efficient Foundation Language Models. *arXiv* **2023**, arXiv:2302.13971. [CrossRef]
5. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. In Proceedings of the 36th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; pp. 27730–27744. [CrossRef]
6. Majumder, S.; Dong, L.; Doudi, F.; Cai, Y.; Tian, C.; Kalathil, D.; Ding, K.; Thatte, A.A.; Li, N.; Xie, L. Exploring the capabilities and limitations of large language models in the electric energy sector. *Joule* **2024**, *8*, 1544–1549. [CrossRef]
7. Shuster, K.; Poff, S.; Chen, M.; Kiela, D.; Weston, J. Retrieval Augmentation Reduces Hallucination in Conversation. In Proceedings of the Findings of the Association for Computational Linguistics, Punta Cana, Dominican Republic, 10 November 2021; pp. 3784–3803. [CrossRef]

8. Farquhar, S.; Kossen, J.; Kuhn, L.; Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature* **2024**, *630*, 625–630. [CrossRef] [PubMed]
9. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; pp. 9459–9474.
10. Zhang, L.; Yu, Y.; Wang, K.; Zhang, C. ARL2: Aligning Retrievers with Black-box Large Language Models via Self-guided Adaptive Relevance Labeling. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 11–16 August 2024; pp. 3708–3719. [CrossRef]
11. Long, Q.; Luo, T.; Wang, W.; Pan, S. Domain Confused Contrastive Learning for Unsupervised Domain Adaptation. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, WA, USA, 10–15 July 2022; pp. 2982–2995. [CrossRef]
12. Azuma, C.; Ito, T.; Shimobaba, T. Adversarial domain adaptation using contrastive learning. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106394. [CrossRef]
13. Ye, D.; Hu, J.; Fan, J.; Tian, B.; Liu, J.; Liang, H.; Ma, J. Best Practices for Distilling Large Language Models into BERT for Web Search Ranking. In Proceedings of the 31st International Conference on Computational Linguistics: Industry Track, Abu Dhabi, United Arab Emirates, 19–24 January 2025; pp. 128–135. [CrossRef]
14. Hong, J.; Tu, Q.; Chen, C.; Xing, G.; Zhang, J.; Yan, R. CycleAlign: Iterative Distillation from Black-box LLM to White-box Models for Better Human Alignment. In Proceedings of the Findings of the Association for Computational Linguistics, Bangkok, Thailand, 11–16 August 2024; pp. 14596–14609. [CrossRef]
15. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. [CrossRef]
16. Fan, W.; Ding, Y.; Ning, L.; Wang, S.; Li, H.; Yin, D.; Chua, T.S.; Li, Q. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Barcelona, Spain, 25–29 August 2024; pp. 6491–6501. [CrossRef]
17. Shao, Z.; Gong, Y.; Shen, Y.; Huang, M.; Duan, N.; Chen, W. Enhancing Retrieval-Augmented Large Language Models with iterative Retrieval-Generation Synergy. In Proceedings of the Findings of the Association for Computational Linguistics, Singapore, 6–10 December 2023; pp. 9248–9274. [CrossRef]
18. Jiang, Z.; Xu, F.; Gao, L.; Sun, Z.; Liu, Q.; Dwivedi-Yu, J.; Yang, Y.; Callan, J.; Neubig, G. Active Retrieval Augmented Generation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 7969–7992. [CrossRef]
19. Shi, W.; Min, S.; Yasunaga, M.; Seo, M.; James, R.; Lewis, M.; Zettlemoyer, L.; Yih, W.t. REPLUG: Retrieval-Augmented Black-Box Language Models. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Mexico City, Mexico, 16–21 June 2024; pp. 8371–8384. [CrossRef]
20. Yoran, O.; Wolfson, T.; Ram, O.; Berant, J. Making Retrieval-Augmented Language Models Robust to Irrelevant Context. *arXiv* **2023**, arXiv:2310.01558. [CrossRef]
21. Li, Y.; Dong, B.; Guerin, F.; Lin, C. Compressing Context to Enhance Inference Efficiency of Large Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 6342–6353. [CrossRef]
22. Kim, J.; Nam, J.; Mo, S.; Park, J.; Lee, S.W.; Seo, M.; Ha, J.W.; Shin, J. SuRe: Summarizing Retrievals using Answer Candidates for Open-domain QA of LLMs. *arXiv* **2024**, arXiv:2404.13081. [CrossRef]
23. Trivedi, H.; Balasubramanian, N.; Khot, T.; Sabharwal, A. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023; pp. 10014–10037. [CrossRef]
24. Gao, J.; Chen, B.; Zhao, X.; Liu, W.; Li, X.; Wang, Y.; Wang, W.; Guo, H.; Tang, R. LLM4Rerank: LLM-based Auto-Reranking Framework for Recommendations. In Proceedings of the ACM on Web Conference, Sydney, NSW, Australia, 28 April–2 May 2025; pp. 228–239. [CrossRef]
25. Hou, Y.; Zhang, J.; Lin, Z.; Lu, H.; Xie, R.; McAuley, J.; Zhao, W.X. Large Language Models are Zero-Shot Rankers for Recommender Systems. In Proceedings of the 46th European Conference on Information Retrieval, Glasgow, UK, 24–28 March 2024; pp. 364–381. [CrossRef]
26. Liu, J.; Ma, Y.; Zhao, R.; Zheng, J.; Ma, Q.; Kang, Y. ListConRanker: A Contrastive Text Reranker with Listwise Encoding. *arXiv* **2025**, arXiv:2501.07111. [CrossRef]
27. Xiong, L.; Xiong, C.; Li, Y.; Tang, K.F.; Liu, J.; Bennett, P.; Ahmed, J.; Overwijk, A. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. *arXiv* **2020**, arXiv:2007.00808. [CrossRef]

28. Tang, H.; Li, H.; Liu, J.; Hong, Y.; Wu, H.; Wang, H. DuReader_robust: A Chinese Dataset Towards Evaluating Robustness and Generalization of Machine Reading Comprehension in Real-World Applications. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021; pp. 955–963. [CrossRef]
29. Lin, C. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Text Summarization Branches Out, Barcelona, Spain, 25–26 July 2004; pp. 74–81.
30. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318. [CrossRef]
31. Jin, J.; Zhu, Y.; Dou, Z.; Dong, G.; Yang, X.; Zhang, C.; Zhao, T.; Yang, Z.; Wen, J.R. FlashRAG: A Modular Toolkit for Efficient Retrieval-Augmented Generation Research. In Proceedings of the ACM on Web Conference, Sydney, NSW, Australia, 28 April–2 May 2025; pp. 737–740. [CrossRef]