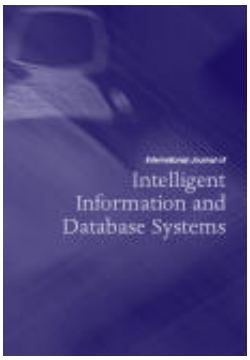
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Intelligent Information and Database Systems

ISSN online: 1751-5866 - ISSN print: 1751-5858
https://www.inderscience.com/ijiids

Adaptive knowledge-driven query generation for zero-shot document retrieval

Weiping Zou, Chenjun Sun, Chao Zheng, Bin Huang

DOI: 10.1504/IJIIDS.2026.10080631

Article History:

Received:	10 August 2025
Last revised:	15 February 2026
Accepted:	29 April 2026
Published online:	04 November 2026

Copyright © 2026 Inderscience Enterprises Ltd.

Adaptive knowledge-driven query generation for zero-shot document retrieval

Weiping Zou*

State Grid Hebei Electric Power Co., Ltd.,
Shijiazhuang City, Hebei Province,
050031, China
Email: weipingzou16@gmail.com
*Corresponding author

Chenjun Sun

State Grid Hebei Electric Power Supply Co., Ltd.,
Information and Telecommunication Branch,
Shijiazhuang City, Hebei Province,
050031, China
Email: chenjuns2022@outlook.com

Chao Zheng and Bin Huang

State Grid Hebei Electric Power Co., Ltd.,
Shijiazhuang City, Hebei Province,
050031, China
Email: zhengchao1033@outlook.com
Email: huangbin198201@outlook.com

Abstract: Document retrieval models often rely on large amounts of labelled data, which are time-consuming and costly to collect and annotate manually. To address this, recent methods generate pseudo queries directly from documents to train retrieval models without manual supervision. However, most existing approaches depend solely on internal document content, overlooking external knowledge that could boost performance, especially in zero-shot settings. In this paper, we propose the adaptive knowledge-driven query expansion with implicit feedback (*ADDEND*), a method that samples representative keywords, expands them using a knowledge graph, and refines pseudo queries to reduce redundancy and semantic drift. Evaluated on 12 publicly available benchmarks, *ADDEND* consistently outperforms existing models, with ablation studies confirming the effectiveness of the keyword sampling and expansion modules.

Keywords: document retrieval; zero-shot learning; knowledge graph; internal documents; external documents.

Reference to this paper should be made as follows: Zou, W., Sun, C., Zheng, C. and Huang, B. (2026) 'Adaptive knowledge-driven query generation for zero-shot document retrieval', *Int. J. Intelligent Information and Database Systems*, Vol. 18, No. 9, pp.1–23.

Copyright © The Author(s) 2026. Published by Inderscience Publishers Ltd. This is an Open Access Article distributed under the CC BY license. (http://creativecommons.org/licenses/by/4.0/)

Biographical notes: Weiping Zou received his Bachelor's degree from Jiangxi University of Finance and Economics, China. He is currently with State Grid Hebei Electric Power Supply Co., Ltd. His research interests include financial accounting, document retrieval and information management, electricity price theory, and economic analysis in power systems.

Chenjun Sun received his Master's degree from North China Electric Power University, China. He is currently with the Information and Telecommunication Branch of State Grid Hebei Electric Power Supply Co., Ltd. His research interests include power grid dispatch, computer applications, information systems, and software technology.

Chao Zheng received his Master's degree from North China Electric Power University, China. He is currently with State Grid Hebei Electric Power Supply Co., Ltd. His research interests include financial accounting, electricity price theory, and management analysis in the electric power industry.

Bin Huang received his Bachelor's degree from North China Electric Power University, China. He is currently with State Grid Hebei Electric Power Supply Co., Ltd. His research interests include financial accounting, electricity price theory, and power enterprise management.

1 Introduction

Document retrieval plays a fundamental role in information systems where the objective is to locate relevant documents from a large corpus in response to a user query. Over the years, document retrieval has evolved significantly, driven by both the increasing complexity of user queries and the growing scale of available data. The ability to accurately retrieve documents has applications across numerous domains, including web search engines (Yao et al., 2024), question answering (Karpukhin et al., 2020), and online recommendation (Yang et al., 2020b).

In its early stages, document retrieval largely relied on statistical and heuristic techniques that focused on matching query terms to document terms. Approaches like term frequency-inverse document frequency (TF-IDF) (Chowdhury, 2010) and BM25 (Robertson et al., 2009) apply sophisticated strategies to weigh the importance of terms in both the query and the documents. These methods assess relevance by measuring how frequently terms appear in a document while simultaneously adjusting for the commonality of the term across the entire corpus. Yet, these conventional methods face several limitations. One key issue is their inability to capture the underlying meaning or semantics of a query. For example, term-based methods often struggle with synonymy (different words with similar meanings) and polysemy (the same word having multiple meanings), leading to suboptimal retrieval results. To overcome these limitations, researchers have increasingly explored dense retrieval methods (Karpukhin et al., 2020; Xiong et al., 2021; Liu et al., 2025) to map textual queries and documents into high-dimensional vector (feature) spaces. Document relevance is then determined by feature similarity between queries and documents. Dense retrieval has demonstrated

significant improvements in various documents retrieval tasks due to its ability to leverage semantic information and contextual relationships.

Despite their success, one major drawback of dense retrieval models is their dependence on large-scale labelled datasets for effective training. Constructing these datasets is often time-consuming and resource-intensive, posing challenges in real-world applications. To address these challenges, zero-shot retrieval techniques have gained significant attention, enabling retrieval in scenarios where labelled data is scarce or even unavailable. In general, zero-shot retrieval models can be categorised into two main classes: *transfer learning-based retrieval approaches* and *generative model-based retrieval approaches*. The former leverages pre-existing data or models and transfers the knowledge learned from the source domain to the target domain (Xin et al., 2022; Xu et al., 2022; Yao et al., 2023); in contrast, the latter utilises generative models to synthesise relevant queries or document representations to facilitate retrieval (Nogueira et al., 2019; Cho et al., 2022; Meng et al., 2024; Engelmann et al., 2024). However, some existing approaches (Nogueira et al., 2019; Cho et al., 2022) lack a feedback mechanism to validate the quality of generated queries, which can result in mismatches with user intent. Additionally, some other methods rely heavily on large-scale language models (such as ChatGPT) (Engelmann et al., 2024), which face limitations in resource demands and may struggle with domain-specific tasks.

To fully harness the contextual information from the input documents and enhance the generated pseudo queries with adaptively sampled keywords that capture core semantics, this paper introduces the adaptive knowledge-driven query expansion with implicit feedback (ADDEND). Our method operates in three key steps:

1 keyword sampling to identify representative keywords from the input document and create a list of preliminary seed keywords

2 keyword expansion to enrich these keywords by leveraging an external knowledge graph, thus covering broader semantic contexts

3 implicit feedback to iteratively refine the query based on both semantic consistency and diversity.

The generated pseudo queries are then used to train the retrieval model using two marginal scopes. Extensive experiments are conducted across 12 highly competitive retrieval benchmarks to demonstrate the effectiveness of ADDEND. Additionally, an ablation study is performed to assess the impact of each component, illustrating how keyword sampling, expansion, and implicit feedback contribute to the model's performance.

The remainder of this paper is organised as follows: Section 2 provides a comprehensive survey of existing research on document retrieval and zero-shot learning for retrieval. Section 3 introduces the proposed method, which involves a keyword sampling/expansion strategy that enhances retrieval effectiveness by leveraging external knowledge and adaptive feedback. Subsequently, in Section 4, the performance of the proposed method is evaluated on 12 highly competitive retrieval benchmarks. Finally, Section 5 concludes this paper by summarising key contributions and suggests future directions for further research in zero-shot retrieval.

2 Related work

2.1 Document retrieval

Document retrieval plays a critical role in many applications, such as search engines (Yao et al., 2024), textual analysis (Yang et al., 2025), and user profiling (Yang et al., 2020a), among others. Traditional retrieval techniques primarily rely on *sparse retrieval* models, such as TF-IDF (Chowdhury, 2010) and BM25 (Robertson et al., 2009), which rank documents based on statistical properties like term frequency and inverse document frequency. Yet, sparse retrieval models struggle to capture *semantic equivalences between words with similar meanings or different forms*, leading to frequent retrieval mismatches and suboptimal outcomes. In contrast, *dense retrieval* techniques leverage semantic representations of textual content in continuous vector spaces (Gao and Callan, 2021; Karpukhin et al., 2020; Xiong et al., 2021), providing a more robust mechanism to address the shortcomings of traditional sparse methods. Specifically, dense retrieval methods are broadly categorised into two primary architectures: the cross-encoder model and the bi-encoder model.

Cross-encoder models employ a joint encoding strategy where both the query and the document are concatenated and passed through a single shared encoder, often a transformer-based architecture (Guu et al., 2020). Let the query be denoted as $\boldsymbol{q}$, with $\boldsymbol{d}^+$ and $\boldsymbol{d}^-$ representing the positive and negative document corresponding to $\boldsymbol{q}$, respectively. The cross-encoder model, denoted as Enc_{cross}, processes concatenated input pairs of the query and documents. Specifically, the concatenated inputs $[\boldsymbol{q}; \boldsymbol{d}^+]$ and $[\boldsymbol{q}; \boldsymbol{d}^-]$ are fed into the model to compute matching or similarity scores for both the positive and negative document. The objective is to ensure that the similarity score for the positive passage, $\text{Enc}_{cross}([\boldsymbol{q}; \boldsymbol{d}^+])$, is maximised, while the similarity score for the negative passage, $\text{Enc}_{cross}([\boldsymbol{q}; \boldsymbol{d}^-])$, is minimised. On the other hand, *bi-encoder models* (Karpukhin et al., 2020) use separate encoders for the query (say Enc_q) and the document (say Enc_d). In this setup, the query and the document are encoded independently into fixed-length vectors. Accordingly, the retrieval process is then simplified to a vector similarity calculation, where the query vector is compared against a pre-computed set of document vectors. This decoupling of encoding allows for significant speedups during inference, as document vectors can be pre-computed and stored in advance.

Despite their success, dense retrieval models typically require large amounts of labelled data for effective training (Zhao et al., 2024). These models rely heavily on supervised learning, where high-quality labelled data is essential to learn dense vector representations of queries and documents. In practice, obtaining such data can be challenging due to the cost and effort involved in manual annotation. Addressing these challenges has become a central research focus, with various methods such as semi-supervised learning (Li et al., 2009), zero/few-shot learning (Cho et al., 2022; Engelmann et al., 2024) emerging as potential solutions. These approaches aim to leverage limited labelled data (semi-supervised and few-shot) or exploit completely unlabelled data (zero-shot) in resource-constrained environments.

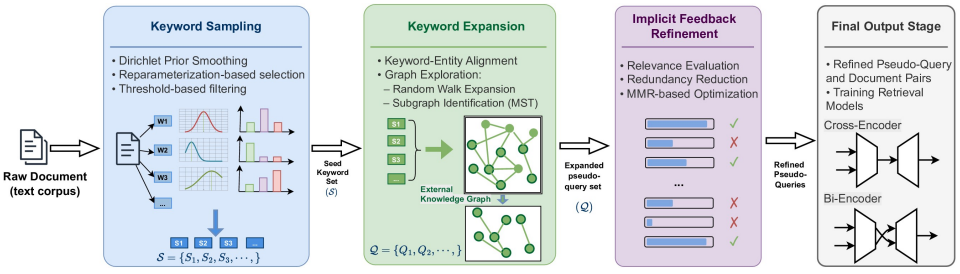
2.2 Zero-shot dense retrieval

The technique of zero-shot dense retrieval has been explored to address the issue of limited labelled data, where models are expected to generalise to unseen data or domains without task-specific supervision. In general, zero-shot dense retrieval models can be categorised into two main classes:

- *Transfer learning-based approaches* leverage pre-existing data or models from a source domain to enable retrieval in a target domain, without relying on task-specific labels (Xin et al., 2022; Xu et al., 2022; Chang et al., 2020; Xiao et al., 2022; Yao et al., 2023). Specifically, in *source domain training*, retrieval models are trained on labelled data from a source domain and are then directly applied to the target domain, assuming that the learned knowledge can generalise across different domains. Additionally, in *model-based approaches*, large pre-trained language models (LMs), such as BERT, RoBERTa, or GPT, which have been trained on extensive corpora of general text data, are commonly used. These LMs can be directly applied to retrieval tasks without requiring task-specific annotations. While these models are capable of understanding general language patterns, they may struggle to capture domain-specific knowledge, which can lead to suboptimal performance in specialised domains.
- *Generative model-based approaches* synthesise pseudo queries from documents to train retrieval models. *docT5query* (Nogueira et al., 2019) first introduced this idea, using a sequence-to-sequence model to map document terms to queries. *QGEK* (Cho et al., 2022) generates $<S$, R, $O>$ triples, where a query seeks an object O related to subject S via relation R, extracted using a fine-tuned BART (Lewis et al., 2020) model. Other methods include summarisation-driven query generation (Meng et al., 2024), GPT-based prompting using document fields (e.g., *title*, *description*, *narrative*) (Engelmann et al., 2024), and human-in-the-loop frameworks such as *QueryExplorer* (Dhole et al., 2024), which refine queries through feedback. *Promptagator* (Dai et al., 2023) prompts ChatGPT with: "($\boldsymbol{d}$) Read the passage and generate a query" for zero-shot query synthesis, where ($\boldsymbol{d}$) represents the document text. *Generate-pairwise* (Chaudhary et al., 2024) further enhances contrastiveness by jointly generating relevant and irrelevant queries conditioned on the same context.

Despite their success, certain approaches (Nogueira et al., 2019; Cho et al., 2022; Meng et al., 2024) lack a feedback mechanism, limiting their ability to validate the quality of generated queries, leading to potential mismatches with user intent. Some retrieval methods rely heavily on large-scale language models (such as ChatGPT) that have limitations in terms of computational resource demands and potential over-reliance on pre-trained distributions, which may not adapt well to domain-specific retrieval tasks (Engelmann et al., 2024). Other methods will require human annotation (Dhole et al., 2024), or extra synthetic data (Dai et al., 2023) or the training of an additional discriminator model (Chaudhary et al., 2024) for validation, which not only increases computational costs but also introduces complexities that may undermine the efficiency and scalability of the retrieval process.

Figure 1 Overview of the proposed the adaptive knowledge-driven query expansion with implicit feedback (ADDEND) algorithm (see online version for colours)



Note: The method initially samples important keywords from the document, then uses a knowledge graph to expand these keywords into semantically related terms (as expanded pseudo query). Those candidate queries are then refined using metrics of query redundancy and semantic mismatch before fine-tuning the final retrieval model.

3 Proposed method

This paper proposes the adaptive knowledge-driven query expansion with implicit feedback (*ADDEND*) algorithm. As illustrated in Figure 1, ADDEND integrates three core modules:

1 *keyword sampling*, utilising the Dirichlet prior smoothing and reparameterisation techniques to extract representative terms from the input document

2 *keyword expansion*, which employs an existing knowledge graph to semantically enrich the sampled keywords, forming the base for initial pseudo-queries

3 *implicit feedback*, which applies both semantic consistency and diversity losses to iteratively refine queries, thereby enhancing retrieval effectiveness and robustness.

3.1 Problem formulation

We consider the problem of zero-shot document retrieval, where no labelled training data is available for target query-document pairs. Let $D = \{\boldsymbol{d}_i\}_{i=1}^{N}$ represent a corpus consisting of N documents. For each document $\boldsymbol{d}_i$, the goal is to generate some corresponding pseudo queries (such as $\boldsymbol{q}_i$) that capture the semantic intent of the document.

Under the zero-shot setting, the retrieval model is trained using such pseudo query-document pairs $(\boldsymbol{q}_i, \boldsymbol{d}_i)$ without explicitly relying on any supervised annotations. As a result, the fine-tuned retrieval model learns to accurately estimate query-document relevance, leading to stable retrieval performance on unseen queries.

3.2 Keyword sampling

In this section, we describe the implementation details for identifying potential keywords, which serve as the foundation for subsequent pseudo-query generation. First,

we leverage the Dirichlet prior smoothing approach to estimate the importance of each word w in a given document $\boldsymbol{d}$. This approach refines the probability distribution of words by smoothing rare terms while preserving the influence of frequently occurring ones, ensuring a more robust keyword selection process. Next, a reparameterisation technique is applied to transform the sampling of keywords into a differentiable process. This allows the model to optimise the keyword selection in an end-to-end manner.

The Dirichlet prior smoothing method (Ma et al., 2021; Zhai and Lafferty, 2017) is widely used for regularising statistical language models. In this work, we adopt this smoothing technique to estimate the importance of each word in a document while accounting for variations in term frequency. Following Dirichlet prior smoothing, the probability of observing word w in document $\boldsymbol{d}$ is defined as:

$$prob(w|\boldsymbol{d}) = \frac{c(w, \boldsymbol{d}) + \mu \times prob(w|D)}{|\boldsymbol{d}| + \mu}, \tag{1}$$

where $c(w, \boldsymbol{d})$ represents the frequency of word w in $\boldsymbol{d}$, $|\boldsymbol{d}|$ denotes the length of the document, i.e., the total number of words in $\boldsymbol{d}$, $prob(w|D)$ refers to the probability of word w being drawn from the overall corpus D, which is proportional to the frequency of w in D, and μ is a smoothing hyper-parameter. With Dirichlet smoothing, word importance is estimated by jointly considering document-specific and corpus-level statistics, enabling the selection of representative keywords.

Once the word importance score $prob(w|\boldsymbol{d})$ is computed, we apply a reparameterisation technique (Jang et al., 2017) to transform these probabilities into an intermediate representation that reflects the likelihood of each word w being selected for expansion later. The reparameterisation process involves applying a deterministic function to $prob(w|\boldsymbol{d})$ as follows:

$$\rho_w = \sigma\left(\frac{\log(\lambda) - \log(1-\lambda) + prob(w|\boldsymbol{d})}{0.2}\right), \tag{2}$$

where $\lambda \sim \text{Uniform}(0, 1)$, and $\sigma(\cdot)$ represents the sigmoid activation function. As such, words with higher importance, as indicated by their $prob(w|\boldsymbol{d})$, are more likely to be selected for subsequent expansion. This selective mechanism prioritises the most contextually relevant words during pseudo-query formation, enhancing the quality and relevance of the retrieval process. To further refine this selection, we introduce a threshold hyperparameter τ_w. That is, only words for which the calculated importance score ρ_w exceeds this threshold, i.e., $\rho_w > \tau_w$, are selected for the next phase of expansion. By integrating this step into an end-to-end learning manner, we ensure that the pseudo-query generation process is both efficient and dynamically adaptable to the content of individual document.

3.3 Keyword expansion

Given the initial set S of keywords extracted from the keyword sampling module, we further enhance these keywords by performing keyword expansion, incorporating external knowledge from a knowledge graph G. The primary objective is to align or match the identified keywords from S to corresponding entities in G and then expand them to enrich the original keyword set with semantically related terms. This expansion

aims to capture additional terms and concepts that may not be explicitly present in the initial document but are contextually relevant, thus broadening the scope of the subsequent pseudo queries. The overall procedure is provided in Algorithm 1.

Initially, the keyword alignment is achieved through either straightforward string matching or more sophisticated techniques (Algorithm 1, lines 1–10). For string matching, keywords from S are directly compared to entity labels in G. However, this approach may overlook semantically relevant entities due to variations in phrasing or terminology across different domains. To address this limitation, we leverage contextual embedding models (e.g., BERT or its variants) to generate dense vector representations (embeddings) for both the keywords and entities in the knowledge graph (Algorithm 1, lines 3–6). These embeddings capture the contextual meaning of each term, allowing the model to align semantically similar entities in G even when exact string matches are absent. The similarity between a keyword's embedding and an entity's embedding is then computed using cosine similarity. A hyperparameter τ_k is introduced as a threshold to determine if the similarity score is sufficiently high for the entity to be considered a valid expansion candidate (Algorithm 1, lines 7–9).

Algorithm 1 Keyword expansion module for pseudo-query generation

Require: Set of initial keywords S, knowledge graph G, similarity threshold τ_k, random walk depth d_w, random subgraph nodes d_n, and pre-trained embedding model M
Ensure: Expanded pseudo-query set Q_{pseudo}
1: **Step 1: Keyword-item matching**
2: **for** each keyword $k \in S$ **do**
3: Generate embedding e_k using pre-trained embedding model M
4: **for** each entity $e \in G$ **do**
5: Generate embedding e_e for entity e using the same model M
6: Compute similarity score $\text{sim}(e_k, e_e)$ using cosine similarity
7: **if** $\text{sim}(e_k, e_e) \geq \tau_k$ **then**
8: Add entity e to the set of valid expansion candidates C_k for keyword k
9: **end if**
10: **end for**
11: **end for**
12: **Step 2: Keyword expansion process**
13: **for** each candidate entity $e_k \in C_k$ aligned with keyword k **do**
14: {%Random walks}
15: Perform a random walk starting from entity e_k with depth d_w in G
16: Collect neighbouring entities/relationship encountered during the walk as expansion
17: {%Subgraph identification}
18: Perform a random selection from C_k to obtain d_n entities
19: Extract a subgraph G_s that connects these d_n entities by constructing a minimum spanning tree (MST)
20: Concatenate all selected entities/relationship to form a structured pseudo-query q_{pseudo}
21: Add q_{pseudo} to the final set Q_{pseudo}
22: **end for**
23: **Return:** S_{pseudo}

Once keywords are aligned with entities in G, we proceed with the keyword expansion process by exploring surrounding entities and their relationships within G and form the pseudo-query (Algorithm 1, lines 12–21). This step aims to enrich the original keyword set with semantically related terms. Two approaches for this exploration are employed:

- *Random walks:* this approach initiates a stochastic traversal of the knowledge graph G, starting from the initially aligned entities. The random walk algorithm explores the graph by visiting neighbouring nodes (entities) based on an equal probability distribution. To ensure that exploration remains focused on the relevant context, a maximal walk depth d_w is set, preventing the traversal from straying too far from the original entities. This process corresponds to Algorithm 1, lines 14–16.
- *Subgraph identification:* this approach aims to identify a subgraph within G that includes as many aligned entities as possible, while maintaining semantic coherence. In random walks, the expansion is based on a single selected starting entity; in subgraph identification, a set of candidate entities is given. By extracting a dense subgraph that maximises the inclusion of these candidate entities, we ensure that the resulting subgraph captures the most relevant relationships and concepts related to the original keywords. The subgraph is constructed using the minimum spanning tree (MST) algorithm, ensuring that only the most contextually significant entities are retained. This step is implemented in Algorithm 1, lines 17–19.

The entities encountered through the above expansion techniques, along with their contextual relationships to the original keywords, are incorporated into the expanded keyword set. Each keyword may map to one or more entities within the knowledge graph G. Additionally, the expansion process could further enable individual keywords to generate multiple expanded keyword sets, each reflecting different facets of its semantic neighbourhood.

The expanded entities are then organised into a structured pseudo-query by concatenating them into a single sequence (Algorithm 1, lines 20–21). The sequence order of these expanded entities is determined based on their proximity to the original keywords in the knowledge graph, reflecting the distance of their connections. This structured pseudo-query not only incorporates the keywords directly associated with the original document but also includes contextually related concepts derived from the knowledge graph. This enriched query is subsequently used as input for fine-tuning the document retrieval model, enhancing the retrieval process by capturing both explicit and implicit semantic relationships.

3.4 Implicit feedback

The pseudo-queries generated by the aforementioned keyword sampling and expansion module, while effective in capturing the diverse semantic landscape of individual documents, may introduce noise that degrades retrieval performance. This noise primarily stems from two key issues:

1 *query redundancy*, where certain pseudo-queries may contain overlapping or repeated information, leading to unnecessary duplication

2 *semantic mismatch*, where the expanded keywords from the knowledge graph introduce terms that, while related, diverge from the core content of the original document.

These issues can significantly impact the quality of the generated pseudo-queries, ultimately affecting the retrieval model effectiveness in retrieving relevant documents.

To address these challenges, we further propose to integrate the maximal marginal relevance (MMR) as a loss or regularisation function during pseudo-query generation. MMR, commonly used to reduce redundancy and improve relevance in information retrieval tasks, helps guide the model toward generating pseudo-queries that maximise both diversity and relevance. In other words, the MMR-based loss is designed to penalise pseudo-queries that are either overly similar to each other or deviate significantly from the original document content. Formally, given the input document $\boldsymbol{d}$ and the relevant set of generated pseudo-queries $Q = \{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n\}$, the MMR loss is defined as follows:

$$\mathcal{L}_{\text{MMR}} = \sum_{\boldsymbol{q}_i \in Q} \Big[\alpha \cdot \big(1 - \text{Relevance}(\boldsymbol{q}_i, \boldsymbol{d})\big) + (1 - \alpha) \cdot \text{Redundancy}(\boldsymbol{q}_i, Q)\Big], \quad (3)$$

where $\text{Relevance}(\boldsymbol{q}_i, \boldsymbol{d})$ measures the semantic coherence between the pseudo-query $\boldsymbol{q}_i$ and the original document $\boldsymbol{d}$, $\text{Redundancy}(\boldsymbol{q}_i, Q)$ evaluates the overlap between the pseudo-query $\boldsymbol{q}_i$ and the other pseudo-queries in the set Q, and α is a hyperparameter that balances the trade-off between maximising relevance and minimising redundancy. Specifically, in this paper, $\text{Relevance}(\boldsymbol{q}_i, \boldsymbol{d})$ is implemented using cosine similarity between the embeddings of the pseudo-query and the document. For $\text{Redundancy}(\boldsymbol{q}_i, Q)$, we compute the pairwise cosine similarity between the embeddings of $\boldsymbol{q}_i$ and the rest of the pseudo-queries in Q.

3.5 Overall model

We integrate the aforementioned three modules (i.e., sampling, expansion, and implicit feedback) into a unified end-to-end framework, enabling automated pseudo-query generation and retrieval model optimisation. The advantage of this end-to-end approach is that it eliminates the need for manual intervention between stages, allowing the model to jointly learn and adjust each module based on the final query-generation performance. This enhances both the efficiency and effectiveness of the retrieval system, as all components are optimised simultaneously, ensuring a more robust zero-shot retrieval capability.

The generated pseudo query-document pairs (denoted as $\boldsymbol{q}$ and $\boldsymbol{d}^+$) are then used to train the retrieval model. The training process consists of three key stages. First, negative passage samples are obtained to provide contrast during training. Specifically, the negative document $\boldsymbol{d}^-$ is retrieved using a pre-trained dense retrieval model. Seven best-match documents (excluding $\boldsymbol{d}^+$) are selected as the negative samples. Next, two marginal scores are computed using both cross- and bi-encoder models, as discussed in Section 2. Specifically, the first marginal score, δ^c, measures the difference between the similarity of $\boldsymbol{q}$ with the positive document $\boldsymbol{d}^+$ and the negative document $\boldsymbol{d}^-$ as:

$$\delta^c = f(\text{Enc}_{cross}(\boldsymbol{q}, \boldsymbol{d}^+)) - f(\text{Enc}_{cross}(\boldsymbol{q}, \boldsymbol{d}^-), \quad (4)$$

where $\text{Enc}_{cross}(\cdot)$ denotes the cross-encoder and $f(\cdot)$ is the similarity scoring function. Similarly, the second marginal score, δ^b, is derived from the bi-encoder model:

$$\delta^b = \text{Enc}_q(\boldsymbol{q})^T \text{Enc}_d(\boldsymbol{d}^+) - \text{Enc}_q(\boldsymbol{q})^T \text{Enc}_d(\boldsymbol{d}^-), \quad (5)$$

where $\text{Enc}_q(\cdot)$ and $\text{Enc}_d(\cdot)$ denote the query and document encoders, respectively.

The final training objective then minimises the discrepancy using both marginal scores from equation (4) and (5) as follows:

$$\mathcal{L}_{retrieval} = -\frac{1}{M}\sum_{i=0}^{M-1}\left|\delta_i^c - \delta_i^b\right|^2, \tag{6}$$

where M is the total number of training examples (or query-document pairs), and δ_i^c and δ_i^b are the marginal scores for the i^{th} training pair. The proposed loss function from equation (6) enforces consistency between the cross- and bi-encoder models, improving the robustness of the retrieval system.

We further investigate the computational complexity for the pseudo-query generation process. The time complexity of ADDEND consists of three main steps:

1 Keyword sampling: for each document, keywords are sampled using Dirichlet prior smoothing and reparameterisation techniques. The Dirichlet prior smoothing can be precomputed, reducing its real-time overhead. Reparameterisation, on the other hand, has a complexity of $\mathcal{O}(n^2)$, where n is the number of sampled keywords per document.

2 Keyword expansion: expanding each sampled keyword involves matching related terms within a knowledge graph. Let d represent the hidden dimension of the embeddings. In the worst-case scenario, the lookup/matching operation requires $\mathcal{O}(n \cdot |G| \cdot d)$, where $|G|$ denotes the total number of entities in the knowledge graph G. Additionally, the random walk operation on G to discover related terms has a complexity of $\mathcal{O}(n \cdot |G| \cdot d_w)$, where d_w is the maximum walk depth and we assume an upper bound of $|G|$ for node connections (or a fully connected graph). Using the Kruskal algorithm, the minimal spanning tree (MST) computation requires $\mathcal{O}(|G| \log |G|)$ complexity.

In summary, the computational complexity of ADDEND primarily depends on the knowledge graph size and can be approximated as $\mathcal{O}(n^2 + n \cdot |G| \cdot d + n \cdot |G| \cdot d_w + |G| \log |G|) \approx \mathcal{O}(n \cdot |G| \cdot d_w)$. This analysis suggests that the processing burden is dominated by the document length, the KG-related size and random walk steps.

4 Experiments

4.1 Setup

4.1.1 Datasets

To evaluate the effectiveness of the proposed ADDEND, we select 12 highly competitive benchmarks from BEIR (Thakur et al., 2021), and their statistics are outlined in Table 1. These datasets encompass a wide range of background domains, including fact verification, question answering, and entity linking, providing a comprehensive assessment of our approach in both general and domain-specific settings.

Table 1 Statistics of employed datasets from the BEIR benchmark

Dataset	*Domain*	*# Query*	*# Document*
ArguAna	Misc.	1,406	8,674
Touche-2020	Misc.	49	382,545
FiQA-2018	Finance	648	57,638
Quora	Quora	10,000	522,931
TREC-COVID	Bio-medical	50	171,332
NFCorpus	Bio-medical	323	3,633
SciFact	Scientific	300	5,183
HotpotQA	Wikipedia	7,405	5,233,329
NQ	Wikipedia	3,452	2,681,468
DBpedia	Wikipedia	400	4,635,922
FEVER	Wikipedia	6,666	5,416,568
Climate-FEVER	Wikipedia	1,535	5,416,593

Note: # Query represents the number of queries and # Document is the total number of documents in the corpus.

4.1.2 Implementation details

For the keyword sampling module, the smoothing parameter $\mu = 0.1$, temperature $T = 0.69$, importance threshold $\tau_w = 0.5$. For the keyword expansion module, we adopt the Wikidata as the knowledge graph. Furthermore, the similarity threshold $\tau_k = 0.6$, the random walk depth $d_w = 5$, the subgraph nodes $d_n = 7$, and the pre-trained embedding model is KG-BERT (Yao et al., 2019). Eventually, each input document is required to produce 5 pseudo queries. For the retrieval model training, the procedure takes the following steps: firstly, a pre-trained dense retrieval model (*ms-marco-distilbert-base-v3*) is used to search for negative samples. This results in the creation of triplets (pseudo query, positive sample, negative sample), which later are employed as the training samples. At last, the *ms-marco-MiniLM-L-6-v2* and *ms-marco-MiniLM-L-6-v3* model is utilised as the cross and bi-encoders to compute the marginal scores between documents and queries.

4.1.3 Evaluation metrics

The normalised discounted cumulative gain at 10 (NDCG@10) is employed in this paper as a standard measure to evaluate retrieval performance. NDCG@10 quantifies the effectiveness of the ranked retrieval results by considering the position and relevance of retrieved documents, where a higher score indicates a more relevant ranking. The NDCG@10 is calculated as follows:

$$NDCG@10 = \sum_{i=1}^{10} \frac{2^{\text{rel}_i} - 1}{\log_2(i+1)} \Big/ \sum_{i=1}^{|REL|} \frac{2^{\text{rel}_i} - 1}{\log_2(i+1)}, \tag{7}$$

where $|REL|$ represents the total length of the candidate list containing retrieved documents, and rel_i denotes the relevance of the i^{th} document.

4.2 Overall result

Baselines: to validate the efficacy of the proposed ADDEND, we compare it with several well-established zero-shot retrieval baseline methods. Specifically, the following methods are employed:

- *RetroMAE* (Xiao et al., 2022): this method utilises a masked autoencoder for retrieval-oriented pre-training. Specifically, the query-document pairs from MS-MARCO (Nguyen et al., 2016) serve as the source domain to pre-train RetroMAE before applying it to the target source.
- *LaPraDoR* (Xu et al., 2022): this method introduces iterative contrastive learning (ICoL) and a lexicon-enhanced dense retrieval process. ICoL incrementally trains both query and document encoders using a cache mechanism, while lexical matching is integrated into the retrieval to improve performance.
- *QGEK* (Cho et al., 2022): this method fine-tunes the BART model (Lewis et al., 2020) to extract knowledge triplets from documents, which are subsequently converted into pseudo queries to support retrieval tasks.
- *Promptagator* (Dai et al., 2023): this approach inputs the original document as a prompt into ChatGPT to generate relevant pseudo queries, leveraging generative language models to enhance retrieval quality.
- *AugTriever* (Meng et al., 2024): this method uses text generation models, originally designed for summarisation tasks, to produce pseudo queries as a form of distant supervision.

The results from the baselines are directly sourced from their original paper. The proposed method reports the average results over five independent runs, each initialised with a distinct random seed.

Table 2 presents the experimental results of the proposed ADDEND compared to existing methods on the zero-shot retrieval task. Specifically, ADDEND achieves the highest performance on 8 out of the 12 benchmark datasets, demonstrating its effectiveness in capturing relevant document features without prior training on these datasets. On the remaining 4 datasets, it secures the second-highest performance, highlighting its robust adaptability across diverse data sources. In terms of the average performance across all 12 datasets, ADDEND consistently outperforms other methods (with the 0.56 NDCG@10), indicating its superior capability in zero-shot retrieval scenarios.

The key advantage of ADDEND lies in its multi-step approach, which leverages keyword sampling, expansion using a knowledge graph, and pseudo query generation to ensure comprehensive coverage and relevance. Compared to *RetroMAE*, which relies on masked autoencoder pre-training, ADDEND shows improvements on several datasets, such as TREC-COVID and Quora, where domain-specific knowledge embedded in keywords plays a significant role in retrieval accuracy. This enhanced performance can be attributed to ADDEND's ability to incorporate domain-relevant expansions through knowledge graphs, which enhances the semantic alignment between queries and documents. *LaPraDoR* utilises lexicon-enhanced dense retrieval to achieve a strong (second-best on average) performance that highlights the importance of lexical matching in downstream retrieval tasks. Our approach similarly incorporates lexical matching

in the knowledge graph expansion stage, where original seed keywords are mapped to entities within the graph. However, ADDEND goes further by integrating semantic matching, leading to enhanced retrieval outcomes by combining both lexical and semantic insights. In contrast to methods like *QGEK* and *Promptagator*, which focus on generating pseudo queries through pre-defined linguistic templates or pre-trained models, ADDEND's knowledge graph-driven expansion is inherently more adaptable to out-of-domain data. Furthermore, while *AugTriever* leverages text generation models for pseudo query generation, it is predominantly tailored for summarisation, leading to over-generalisation in retrieval contexts. ADDEND, with its knowledge graph-guided expansion, enables targeted and domain-specific formation, which improves the retrieval performance significantly.

Table 2 Performance comparison (NDCG@10) on the proposed ADDEND and existing methods for zero-shot retrieval using the BEIR benchmark

Dataset	*RetroMAE*	*LaPraDoR*	*QGEK*	*Promptagator*	*AugTriever*	*ADDEND*
ArguAna	0.433	0.508	0.095	**0.538**	0.398	*0.523*$^{\blacklozenge}$
Touche-2020	0.237	*0.333*	0.209	0.266	0.198	**0.396**$^{\blacklozenge}$
FiQA-2018	0.316	0.343	0.118	**0.404**	0.292	*0.344*$^{\blacklozenge}$
Quora	0.847	*0.875*	0.498	–	0.799	**0.896**$^{\blacklozenge}$
TREC-COVID	0.772	*0.779*	0.441	0.727	0.640	**0.824**$^{\blacklozenge}$
NFCorpus	0.308	0.347	0.179	0.334	*0.351*	**0.378**$^{\blacklozenge}$
SciFact	0.653	*0.687*	0.344	0.623	*0.687*	**0.695**$^{\blacklozenge}$
HotpotQA	0.635	**0.666**	0.423	0.604	0.576	*0.641*$^{\blacklozenge}$
NQ	**0.518**	0.479	0.284	–	0.335	*0.501*$^{\blacklozenge}$
DBpedia	0.390	*0.391*	0.273	0.364	0.327	**0.410**$^{\blacklozenge}$
FEVER	*0.774*	0.763	0.773	0.762	0.653	**0.799**$^{\blacklozenge}$
Climate-FEVER	0.232	*0.261*	0.175	0.214	0.159	**0.306**$^{\blacklozenge}$
Avg.	0.510	*0.536*	0.318	0.484	0.451	**0.560**

Notes: The best score on a given dataset is marked in **bold**, while the second best is *italic*. $^{\blacklozenge}$ indicates significant at $p < 0.05$ and $-$ indicates no relevant results available.

4.3 Ablation study

To provide a comprehensive analysis and deepen our understanding of the effectiveness of the proposed ADDEND, we conduct experiments designed to address the following research questions:

- *Effectiveness of keyword sampling:* how effectively does ADDEND select representative keywords from documents to generate high-quality pseudo-queries for zero-shot retrieval?
- *Impact of knowledge graph expansion:* how does the integration of knowledge graphs for keyword expansion influence retrieval performance in zero-shot settings?
- *Model component analysis:* how do the individual components of ADDEND, such as keyword sampling and knowledge graph expansion, contribute to the overall

retrieval performance? This analysis aims to understand the marginal effect of each component by isolating them and evaluating their performance contributions independently.

- *Hyperparameter sensitivity analysis:* what is the impact of hyperparameters on the performance of ADDEND? In particular, we examine the influence of the importance threshold τ_w, which is used to select candidate seed keywords, and the expansion threshold τ_k, which is applied to match keywords with entities from the knowledge graph. Additionally, what is the optimal value for α?

To evaluate ADDEND across diverse domains, we select the Quora, NFCorpus, and FEVER datasets, representing varied document types and sizes within the BEIR benchmark. All experiments are conducted five times, and the average results are reported to ensure statistical reliability.

4.3.1 Effectiveness of keyword sampling

Our proposed algorithm initiates by extracting essential keywords from the original documents, employing Dirichlet prior smoothing combined with a reparameterisation technique to enhance keyword selection flexibility. To evaluate the effectiveness of our keyword-sampling strategy, we introduce the following variants for comparison:

1 *POS*, where only nouns, verbs, adjectives, and adverbs are sampled randomly from the text

2 *TFIDF*, which selects keywords based on high TF-IDF scores, emphasising frequently relevant terms

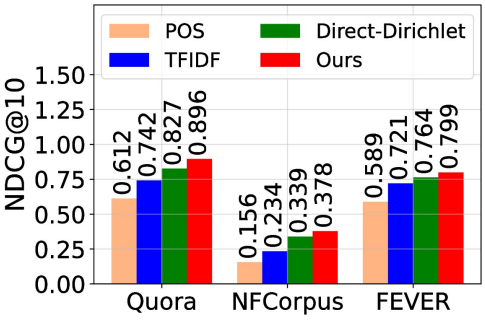
3 *Direct-Dirichlet*, where keywords are extracted directly through Dirichlet prior smoothing without applying the reparameterisation trick.

Each variant allows us to systematically assess the impact of different keyword-sampling strategies on retrieval performance.

The retrieval results are illustrated in Figure 2, where our proposed method exhibits a significant performance improvement. The *POS* approach yields the lowest performance, likely due to its reliance on random sampling, which can include terms with limited relevance to the document main topics. This lack of specificity reduces the effectiveness in representing essential concepts for retrieval. The *TFIDF* variant improves upon POS by prioritising frequently relevant terms, yet high TF-IDF scores can introduce redundancy, potentially neglecting less frequent but contextually vital terms, thereby limiting overall retrieval quality. The *Direct-Dirichlet* method demonstrates notable improvement, leveraging document likelihoods for keyword weighting to achieve a more stable selection. However, the absence of reparameterisation in this variant reduces its adaptability, resulting in keyword sets that may not adequately capture nuanced document contexts. In contrast, our proposed method integrates both the reparameterisation trick and knowledge-driven keyword expansion, offering a dynamic approach to keyword selection. The reparameterisation trick enhances adaptability by avoiding excessive reliance on overly probable terms that may not align with broader contextual significance. When combined with knowledge graph expansion, our method captures richer semantic relationships, enabling the retrieval model to achieve superior

performance by incorporating both contextually relevant and semantically diverse keywords.

Figure 2 Impact of different keyword sampling strategies on the subsequent retrieval performance (NDCG@10) (see online version for colours)



4.3.2 *Impact of knowledge graph expansion*

Our proposed algorithm leverages keyword expansion techniques grounded in random walks and subgraph identification to enhance pseudo-queries. To systematically assess the effectiveness of these techniques, we introduce the following variants:

1 *Neighbour-only expansion (NOE):* this variant selects only the direct neighbours of each keyword entity from the knowledge graph, focusing on immediate and highly relevant concepts closely related to the keywords.

2 *Random walk-only expansion (RWE):* this variant utilises random walk-based sampling to explore the knowledge graph, capturing semantically similar but not immediately adjacent keywords. The random walk enables a broader search within the graph, potentially identifying indirect yet contextually meaningful terms.

3 *Subgraph-only expansion (SGE):* this variant performs subgraph identification to uncover clusters of interrelated keywords, incorporating a more extensive set of interconnected terms beyond the random-walk one.

This method allows for a richer, contextually expansive keyword set that enhances the depth of pseudo-queries. These variants are designed to exploit distinct aspects of the knowledge graph, creating keyword expansions that support effective pseudo-query generation. Table 3 presents the experimental results for each variant over five runs, evaluating their performance across multiple datasets.

Table 3 Performance comparison (NDCG@10) of keyword expansion variants on zero-shot document retrieval

Variant	*Quora*	*NFCorpus*	*FEVER*
NOE	0.764	0.310	0.672
RWE	0.820	0.345	0.735
SGE	0.860	0.370	0.770
Full	0.896	0.378	0.799

The results demonstrate distinct contributions from each keyword expansion method to retrieval performance. The *NOE* variant yields the lowest NDCG@10 scores, performing less effectively across all datasets. This variant, by focusing on direct neighbours, captures only highly relevant yet limited terms. Consequently, its performance suffers on datasets like NFCorpus and FEVER, which benefit from a broader contextual understanding. The *RWE* variant, which employs random walk sampling, achieves improved performance over NOE by 5.6% on Quora, 3.5% on NFCorpus, and 9.3% on FEVER. This improvement likely stems from the inclusion of semantically similar terms from broader graph explorations, enriching the keyword set with indirect but contextually relevant words. *SGE* further boosts retrieval effectiveness, particularly on Quora and FEVER, with enhancements over NOE of 12.6% and 14.5%, respectively. By identifying clusters of related keywords, SGE provides a more comprehensive term set, effectively capturing the document's contextual background and resulting in higher retrieval accuracy. At last, our proposed method surpasses all three variants across datasets, delivering an average improvement of 4.2% over SGE. This performance advantage highlights the effectiveness of our approach, which combines both contextual expansion and clustering information. The resulting pseudo-queries are not only semantically diverse but also closely aligned with relevant document contexts, thus optimising retrieval outcomes in zero-shot scenarios.

4.3.3 Model breakdown

In this section, we conduct an ablation study to systematically analyse the contribution of each component within ADDEND, with a particular focus on *keyword sampling* (KS) and *knowledge graph expansion* (KGE). By evaluating retrieval performance when each component is included or excluded, we aim to gain a clearer understanding of their roles in retrieval and improving robustness in diverse contexts. To assess the impact of each component individually, we define the following model variants:

1 *Full:* the complete model, integrating both keyword sampling and knowledge graph expansion. This configuration represents ADDEND's full potential in leveraging sampled keywords and external knowledge, thus serving as the baseline for comparison.

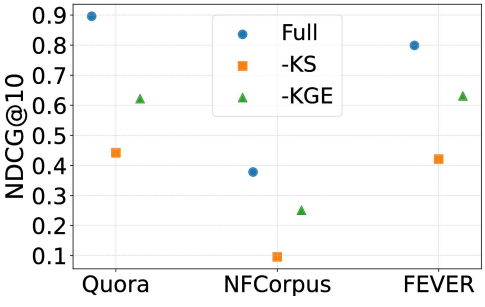
2 *-KS (without keyword sampling):* this variant omits keyword sampling, instead directly selecting keywords with the highest TF-IDF values from the document. This approach tests the effectiveness of keyword sampling in identifying representative and contextually relevant terms beyond simple frequency-based selection.

3 *-KGE (without knowledge graph expansion):* this variant excludes knowledge graph expansion, limiting the pseudo-query to keywords sampled directly from the document without external knowledge augmentation. Here, we investigate the significance of using a knowledge graph for expanding the retrieval context and connecting semantically related terms.

As observed in Figure 3, the *Full* model consistently outperforms both the *-KS* (without keyword sampling) and *-KGE* (without keyword graph expansion) configurations across all evaluated datasets. Specifically, for the FEVER dataset, the *Full* model achieves

a score of 0.799, while the -*KS* and -*KGE* variants only reach 0.421 and 0.631, respectively. This performance difference highlights that both keyword sampling and keyword graph expansion contribute significantly to the effectiveness of pseudo-query generation, with keyword sampling showing a notably stronger impact. In particular, the relatively smaller drop for -*KGE* suggests that the model's ability to capture core information from documents relies more heavily on the sampling of relevant keywords. This sampling process likely strengthens the model by focusing on core concepts within a document, enabling pseudo-queries that align closely with the target information. Without keyword sampling, the generated pseudo-queries may lack essential specificity, resulting in lower retrieval accuracy.

Figure 3 Performance comparison (NDCG@10) for the model breakdown on the subsequent retrieval performance (see online version for colours)



Conversely, the inclusion of keyword expansion further enhances retrieval performance by broadening the keyword set with semantically related terms. This expansion increases the semantic coverage of the queries, making them more robust in matching the target content across varied contexts. Thus, while keyword sampling appears essential for maintaining specificity, keyword expansion serves to diversify and deepen the query alignment with the document semantic.

4.3.4 Hyperparameter sensitive analysis

In this section, we firstly investigate the impact of two key hyperparameters on retrieval performance within our zero-shot retrieval framework. Specifically, hyperparameters include:

1 the importance threshold τ_w used in the *keyword sampling* module

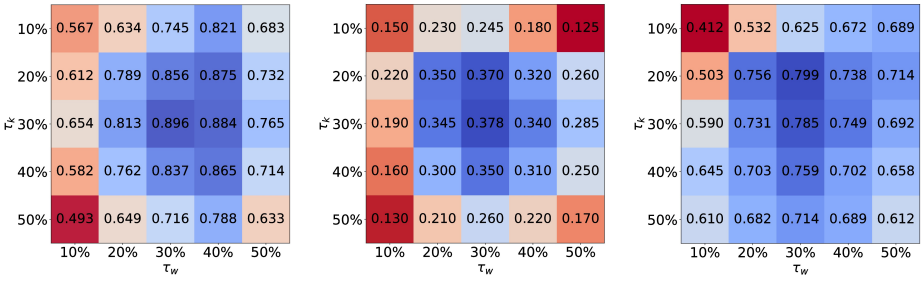
2 the threshold τ_k applied in the *keyword expansion* module for matching keywords to entities within the knowledge graph.

The parameter τ_w is responsible for selecting candidate seed keywords, while τ_k controls the matching sensitivity for expanding these keywords through knowledge graph entities. Both parameters are evaluated in the range of 10% to 50%, with increments of 10%. The average retrieval performance results over five runs are presented in Figure 4.

As illustrated in Figure 4, increasing the threshold τ_w reduces the number of selected keywords in the sampling module, leading to a smaller pool of keywords

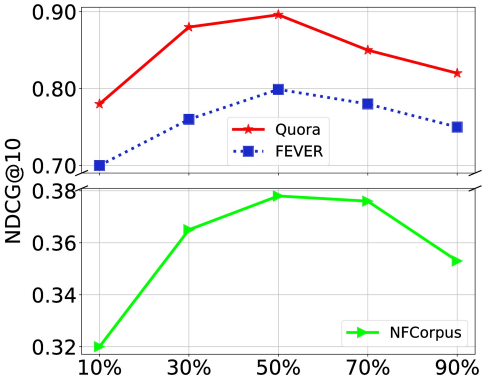
for subsequent expansion. This controlled selection can enhance retrieval precision by limiting the inclusion of irrelevant keywords, thus focusing the expansion process on more representative keywords. However, if τ_w is set too high, certain essential keywords may be omitted, potentially reducing coverage and impacting the retrieval of relevant documents. Conversely, setting τ_w to lower values increases the pool of selected keywords, thereby broadening the keyword base for expansion. While this can enhance recall by covering a wider range of topics, it may also introduce noise due to the inclusion of less relevant or contextually inappropriate keywords.

Figure 4 Retrieval performance measured by NDCG@10 under hyperparameter sensitivity analysis, demonstrating the effects of the importance threshold τ_w (x-axis) and the keyword matching threshold τ_k (y-axis) across different datasets, (a) Quora (b) NFCorpus (c) FEVER (see online version for colours)



Note: Higher values indicate better performance.

Figure 5 Impact analysis from retrieval performance (NDCG@10) comparison on the hyperparameter of α, i.e., the trade-off between relevance and redundancy in the model training loss (see online version for colours)



Similarly, the threshold τ_k directly affects the breadth of matched entities from the knowledge graph for query generation. As τ_k increases, the number of matched entities decreases, restricting the expansion scope to the most relevant entities. This restriction can benefit precision by focusing the expansion on entities that are highly relevant to the original keywords. However, a higher τ_k may narrow the retrieval scope, reducing the model ability to capture diverse document content. In contrast, lowering τ_k expands

the pool of matched entities, potentially improving recall by allowing a broader entity set for query generation. However, this expanded pool may also introduce noise, as less relevant entities are included, potentially diluting the quality of the generated queries. Overall, experimental results reveal a trade-off between the settings of τ_w and τ_k. Higher thresholds tend to limit expansion to the most representative keywords and entities, while lower thresholds favour recall by broadening the keyword and entity selection at the cost of noise. These findings provide insight into optimising τ_w and τ_k to achieve a model balance.

We further examine the role of the hyperparameter α from equation (3), which governs the balance between relevance and redundancy in the training loss. We conduct an evaluation of α with values set from 0.1 to 0.9 in increments of 0.2. The mean retrieval results across five experimental runs for each α value are depicted in Figure 5.

As shown, low values of α (e.g., 0.1) prioritise reducing redundancy by generating a more diverse set of pseudo-queries. While this diversity might lead to unique queries, it can also compromise query relevance, thereby slightly lowering retrieval accuracy. Conversely, high values of α (e.g., 0.9) focus on maximising relevance, producing queries that closely align with the reference passage. This emphasis on alignment, however, introduces high redundancy among queries generated from the same passage, which may diminish the diversity of retrieved results. Intermediate values of α (e.g., between 0.3 and 0.7) appear to provide a balanced trade-off, promoting both relevance and diversity. This balanced setting yields pseudo-queries that enhance overall retrieval performance. For instance, on the FEVER dataset, the model achieves its peak performance with an accuracy of 0.799 at $\alpha = 0.5$. These findings highlight the importance of a balanced approach, where an optimal α value mediates between relevance and redundancy to enhance retrieval accuracy.

5 Conclusions

Zero-shot retrieval aims to retrieve relevant documents without labelled training queries, enabling models to generalise to new domains without costly annotations. This paper introduces ADDEND: adaptive knowledge-driven query expansion with implicit feedback. It improves retrieval by dynamically expanding queries to balance relevance and diversity. ADDEND comprises:

1 keyword sampling, which selects key terms using Dirichlet prior smoothing and reparameterisation

2 keyword expansion, which enriches sampled terms via knowledge graphs

3 implicit feedback, which iteratively refines pseudo-queries using relevance-diversity signals aligned with input documents.

Experiments on 12 public benchmarks show that ADDEND consistently outperforms baselines in zero-shot settings. Future work will explore integrating knowledge graphs with large-scale language models to further enhance query quality and retrieval efficiency.

Acknowledgements

The authors would like to thank anonymous reviewers for their valuable suggestions to improve the quality of the article. This work is supported by State Grid Hebei Electric Power Co., Ltd (China). Science and Technology Project 'Research and Application of Key Technologies for Integrated Internal and External Documents Based on Collaborative Co-Signing and Intelligent Checking Model' (B304XA220062).

Declarations

All authors declare that they have no conflicts of interest.

References

Chang, W-C., Yu, F.X., Chang, Y-W., Yang, Y. and Kumar, S. (2020) 'Pre-training tasks for embedding-based large-scale retrieval', *ICLR*, OpenReview.net.

Chaudhary, A., Raman, K. and Bendersky, M. (2024) 'It's all relative! – A synthetic query generation approach for improving zero-shot relevance prediction', *Findings of the Association for Computational Linguistics: NAACL 2024*, pp.1645–1664.

Cho, S., Jeong, S., Yang, W. and Park, J.C. (2022) 'Query generation with external knowledge for dense retrieval', *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp.22–32.

Chowdhury, G.G. (2010) *Introduction to Modern Information Retrieval*, Facet Publishing, London.

Dai, Z., Zhao, V.Y., Ma, J., Luan, Y., Ni, J., Lu, J., Bakalov, A., Guu, K., Hall, K. and Chang, M-W. (2023) 'Promptagator: few-shot dense retrieval from 8 examples', *The Eleventh International Conference on Learning Representations*.

Dhole, K., Bajaj, S., Chandradevan, R. and Agichtein, E. (2024) 'QueryExplorer: an interactive query generation assistant for search and exploration', *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pp.107–115.

Engelmann, B., Breuer, T., Friese, J.I., Schaer, P. and Fuhr, N. (2024) 'Context-driven interactive query simulations based on generative large language models', *Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Proceedings, Part II*, 24–28 March, Glasgow, UK, pp.173–188.

Gao, L. and Callan, J. (2021) 'Condenser: a pre-training architecture for dense retrieval', *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp.981–993.

Guu, K., Hashimoto, T.B., Oren, Y. and Liang, P. (2020) 'Retrieval-augmented generation for knowledge-intensive NLP tasks', *Proceedings of the 37th International Conference on Machine Learning*, PMLR, pp.3929–3938.

Jang, E., Gu, S. and Poole, B. (2017) 'Categorical reparameterization with Gumbel-softmax', *International Conference on Learning Representations*.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D. and Yih, W-t. (2020) 'Dense passage retrieval for open-domain question answering', *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp.6769–6781.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. and Zettlemoyer, L. (2020) 'BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension', *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp.7871–7880.

Li, M., Li, H. and Zhou, Z-H. (2009) 'Semi-supervised document retrieval', *Inf. Process. Manage.*, Vol. 45, No. 3, pp.341–355.

Liu, J., He, J., Hu, X., Yang, W., Yang, J. and Guo, Y. (2025) 'Importance-awareness masking network for robust document retrieval', *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.1–5.

Ma, X., Guo, J., Zhang, R., Fan, Y., Ji, X. and Cheng, X. (2021) 'Prop: pre-training with representative words prediction for ad-hoc retrieval', *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp.283–291.

Meng, R., Liu, Y., Yavuz, S., Agarwal, D., Tu, L., Yu, N., Zhang, J., Bhat, M. and Zhou, Y. (2024) 'AugTriever: unsupervised dense retrieval by scalable data augmentation', *Proceedings of the 3rd International Workshop on Data-Centric AI at CIKM*.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R. and Deng, L. (2016) 'Ms marco: a human generated machine reading comprehension dataset', *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (CEUR Workshop Proceedings, Vol. 1773)*, CEUR-WS.org.

Nogueira, R., Yang, W., Lin, J. and Cho, K. (2019) *Document Expansion by Query Prediction*, arXiv preprint arXiv:1904.08375.

Robertson, S., Zaragoza, H., et al. (2009) 'The probabilistic relevance framework: Bm25 and beyond', *Foundations and Trends® in Information Retrieval*, Vol. 3, No. 4, pp.333–389.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A. and Gurevych, I. (2021) 'BEIR: a heterogeneous benchmark for zero-shot evaluation of information retrieval models', *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Xiao, S., Liu, Z., Shao, Y. and Cao, Z. (2022) 'RetroMAE: pre-training retrieval-oriented language models via masked auto-encoder', *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp.538–548.

Xin, J., Xiong, C., Srinivasan, A., Sharma, A., Jose, D. and Bennett, P. (2022) 'Zero-shot dense retrieval with momentum adversarial domain invariant representations', *Findings of the Association for Computational Linguistics: ACL 2022*, pp.4008–4020.

Xiong, L., Xiong, C., Li, Y., Tang, K-F., Liu, J., Bennett, P.N., Ahmed, J. and Overwijk, A. (2021) 'Approximate nearest neighbor negative contrastive learning for dense text retrieval', *International Conference on Learning Representations*.

Xu, C., Guo, D., Duan, N. and McAuley, J. (2022) 'LaPraDoR: unsupervised pretrained dense retriever for zero-shot text retrieval', *Findings of the Association for Computational Linguistics: ACL 2022*, pp.3557–3569.

Yang, J., Ma, J. and Howard, S.K. (2020a) 'Usage profiling from mobile applications: A case study of online activity for Australian primary schools', *Knowledge-Based Systems*, Vol. 191, p.105214.

Yang, J., Yecies, B. and Zhong, P.Y. (2020b) 'Characteristics of Chinese online movie reviews and opinion leadership identification', *International Journal of Human-Computer Interaction*, Vol. 36, No. 3, pp.211–226.

Yang, W., Yang, J., Guo, Y. and Barthelemy, J. (2025) 'Defensive dual masking for robust adversarial defense', *Computational Linguistics*, Vol. 52, No. 1, pp.151–190.

Yao, L., Mao, C. and Luo, Y. (2019) *KG-BERT: BERT for Knowledge Graph Completion*, arXiv 1909.03193.

Yao, X., Ma, J., Hu, X., Wang, H., Huang, G., Zhang, L., Yang, J. and Li, W. (2023) 'Killing many birds with one stone: single-source to multiple-target domain adaptation for extractive question answering', *2023 IEEE Smart World Congress (SWC)*, pp.1–8.

Yao, X., Zhang, Z., Hu, X., Yang, J., Guo, Y. and Zhu, D.D. (2024) 'COTER: conditional optimal transport meets table retrieval', *Proceedings of the 17th ACM International Conference on Web Search and Data Mining, WSDM '24*, pp.911–919.

Zhai, C. and Lafferty, J. (2017) 'A study of smoothing methods for language models applied to ad hoc information retrieval', *ACM SIGIR Forum*, ACM New York, NY, USA, Vol. 51, pp.268–276.

Zhao, W.X., Liu, J., Ren, R. and Wen, J-R. (2024) 'Dense text retrieval based on pretrained language models: a survey', *ACM Trans. Inf. Syst.*, Vol. 42, No. 4, pp.1–60.