QAEncoder: Towards Aligned Representation Learning in Question Answering Systems

Zhengren Wang[1,2†], **Qinhan Yu**[1†], **Shida Wei**[1]
Zhiyu Li[2*], **Feiyu Xiong**[2], **Xiaoxing Wang**[2], **Simin Niu**[2], **Hao Liang**[1,3], **Wentao Zhang**[1,2,3*]
[1]Peking University [2]Institute for Advanced Algorithms Research [3]Zhongguancun Academy
wzr@stu.pku.edu.cn {lizy,xiongfy}@iaar.ac.cn wentao.zhang@pku.edu.cn

Abstract

Modern QA systems entail retrieval-augmented generation (RAG) for accurate and trustworthy responses. However, the inherent gap between user queries and relevant documents hinders precise matching. We introduce QAEncoder, a training-free approach to bridge this gap. Specifically, QAEncoder estimates the expectation of potential queries in the embedding space as a robust surrogate for the document embedding, and attaches document fingerprints to effectively distinguish these embeddings. Extensive experiments across diverse datasets, languages, and embedding models confirmed QAEncoder's alignment capability, which offers a simple-yet-effective solution with zero additional index storage, retrieval latency, training costs, or catastrophic forgetting and hallucination issues. The repository is publicly available at https://github.com/IAAR-Shanghai/QAEncoder.

1 Introduction

"What I cannot create, I do not understand." — Richard Feynman

Question Answering (QA) systems aim to generate accurate responses to user queries with applications in customer service (Xu et al., 2024), search engine (Ojokoh and Adebisi, 2018), healthcare (Guo et al., 2022) and education (Levonian et al., 2023). Modern QA systems leverage large language models (LLMs) such as ChatGPT (Achiam et al., 2023), supplemented with retrieval-augmented generation (RAG) to address issues of outdated or hallucinatory information, especially for rapidly evolving knowledge bases (Lewis et al., 2020; Huang et al., 2023; Gupta et al., 2024). The efficacy of RAG hinges on its retrieval module for identifying relevant documents from a vast corpus. Dense retrievers (Lewis et al., 2020; Hofstätter et al., 2021), contrasted with keyword-matching-based sparse retrievers (Jones, 1973; Robertson and Zaragoza, 2009), have enabled more precise retrieval by mapping queries and documents into a shared vector space. Despite advancements, a significant challenge that persists is bridging the gap between user queries and documents across lexical, syntactic, semantic, and content dimensions (Zheng et al., 2020; Nogueira et al., 2019; Cheriton, 2019), termed *the document-query gap*. Three main approaches have emerged to address this challenge: training-based, document-centric and query-centric alignment.

Training-based approaches (Dong et al., 2022; Li et al., 2022; W et al., 2023; Zhang et al., 2024a; Khanna and Subedi, 2024) directly train embedding models with QA datasets to close the representation of relevant queries and documents. But for out-of-domain or multi-domain adaption settings, they struggle to fully generalize (Suprem and Pu, 2022) and face catastrophic forgetting issues. Document-centric approaches (Wang et al., 2023b; Gao et al., 2023; Kim and Min, 2024) generate pseudo-documents for user queries via LLMs, which are then used as retrieval queries; however, this inference process is costly, time-consuming, and prone to hallucinations (Wang et al., 2023b).

In contrast, query-centric methods (Nogueira et al., 2019) generate and index potential queries to better match user queries. Yet, existing methods mainly focus on sparse retrievers and have not fully leveraged the potential of dense retrievers (Cheriton, 2019; Mallia et al., 2021). The naive attempt is to directly store predicted QA pairs into a vector database, but with evident drawbacks like expanded index size, longer retrieval latency, and limited query handling. Therefore, integrating query-centric methods with dense retrievers remains challenging and heavily under-explored.

Motivation Inspired by Feynman's philosophy of learning, in QA systems, effective information

† Equal contribution; * Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4306–4332

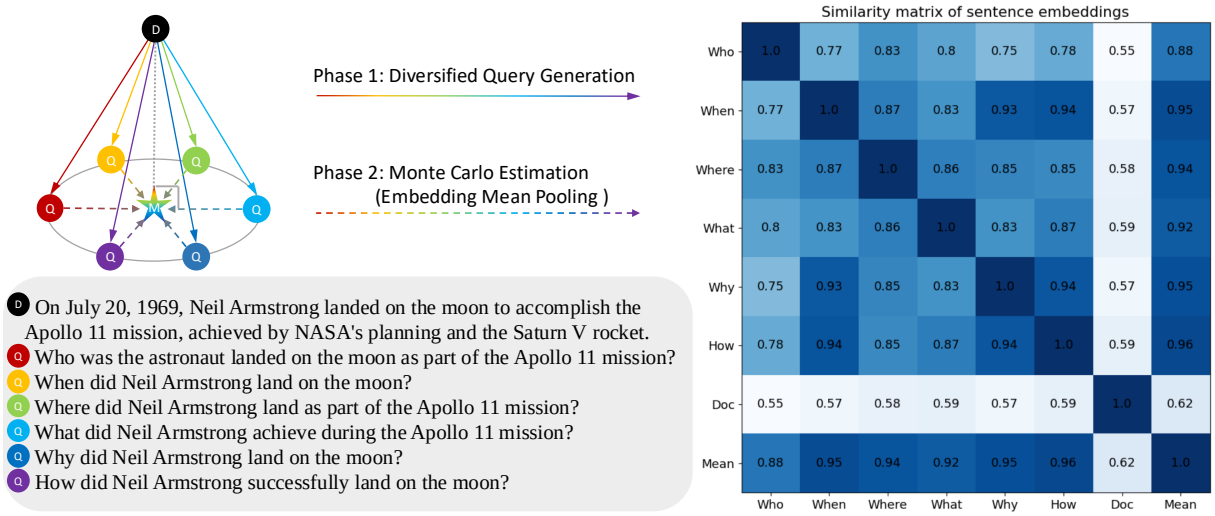


Figure 1: Illustration of QAEncoder's alignment process. Left: Solid lines represent diversified query generation, while dashed lines indicate Monte Carlo estimation. Right: The heatmap depicts the similarity scores among the embeddings of different queries, the document, and the mean estimation. Compared to the document itself, the mean estimation is significantly better aligned with different queries, i.e. a robust surrogate for the document embedding.

retrieval extends beyond mere storage and involves creative processes like query formulation and summarization (Reyes et al., 2021). For instance, the well-established 5W1H framework (Who, What, When, Where, Why, How) help systematically deconstruct information and actively foster a deeper understanding (Jinks, 2019). In this paper, we continue the research line of query-centric methods, and propose *QAEncoder* as a pioneering work.

As demonstrated in Fig. 1, our method initially generates diversified queries (e.g. 5W1H), and then estimates the cluster center of potential queries by the Monte Carlo method. The similarity matrix reveals that, for any query, the mean-query similarity is significantly higher than both document-query and other query-query similarities. Hence, we advocate using the cluster center as a surrogate for the document embedding, which bridges the document-query gap robustly without extra index size and retrieval latency. Meanwhile, we propose the conical distribution hypothesis for more theoretical grounding, which intuitively visualizes the document-query gap and its alignment by abstracting geometric properties of the embedding space.

Despite these advantages, this basic proposal encounters a critical challenge. While enhancing similarity with user queries, it simultaneously reduces the distinguishability between document representations, as they all become query-like. To address this side effect, we further propose *document fingerprint strategies* to reintroduce unique document identities into representations and enable state-of-the-art performance.

Contributions The contributions are threefold:

- **Methodological Innovations.** We pioneer to bridge the document-query gap in dense retrievers from the query-centric perspective. Our method, QAEncoder, not only avoids extra index storage, retrieval latency, training cost and hallucination, but also guarantees diversified query handling and robust generalization. We also propose document fingerprints to address the side effect on distinguishability and achieve state-of-the-art performance.
- **Theoretical Discovery.** We formulate the conical distribution hypothesis, and validate it through empirical analysis. This hypothesis provides not only deeper insights into the geometry of semantic space, but also theoretical foundations for the alignment process.
- **Practical Applications.** QAEncoder is a simple-yet-effective plugin, which seamlessly integrates with existing RAG architectures and training-based methods. This integration significantly boosts system performance with minor modifications required.

2 Related Work

Retrieval-augmented QA systems. Retrieval-augmented generation significantly improves large language models in QA systems by incorporating

a retrieval module that fetches relevant information from external knowledge sources (Févry et al., 2020; Guu et al., 2020; Izacard and Grave, 2021; Zhao et al., 2024). Retrieval models have evolved from early sparse retrievers, such as TF-IDF (Jones, 1973) and BM25 (Robertson and Zaragoza, 2009), which rely on word statistics and inverted indices, to dense retrieval strategies (Lewis et al., 2020) that utilize neural representations for enhanced semantic matching. Advanced methods, such as Self-RAG (Asai et al., 2023) which determines if additional information is required and evaluates the relevance of retrieved content, and RAG-end2end (Siriwardhana et al., 2023) that jointly trains the retriever and generator, represent significant developments in this area. However, these methods still ignore the document-query gap.

Training-based alignment. Training-based approaches bridge the document-query gap generally by contrastive learning (Xiong et al.; Qu et al., 2021) or knowledge distillation (Zhang et al., 2024a; Khanna and Subedi, 2024). For instance, Dong et al. (2022) showed parameter sharing of the query encoder and the document encoder improves overall performance by projecting queries and documents into shared space. Dual-Cross-Encoder (Li et al., 2022) and Query-as-context (W et al., 2023) train embedding models from scratch with paired document-query samples. GPL (Wang et al., 2021), CAI (Iida and Okazaki, 2022) and AugTriever (Meng et al., 2022) conduct domain adaptation with pseudo-queries. However, training-based methods usually face generalization difficulty (Suprem and Pu, 2022), and catastrophic forgetting issues (Pan et al., 2024; Saunders, 2022).

Document-centric alignment. Document-centric methods, such as HyDE (Gao et al., 2023) and Query2doc (Wang et al., 2023b), dynamically transform user queries into pseudo-documents using LLMs for both sparse and dense retrievers. QA-RAG (Kim and Min, 2024) advances by implementing a two-way retrieval mechanism that utilizes both user query and pseudo-documents for respective retrieval and ranking. However, their effectiveness is highly dependent on the quality of pseudo-documents, which are prone to hallucinations, especially for the latest information. Furthermore, during inference, invoking LLMs for user queries imposes extra computational cost and latency, leading to degraded user experience.

Query-centric alignment. The seminal work, Doc2Query (Nogueira et al., 2019), focuses on the vocabulary mismatch problem for sparse retrievers by expanding the document with keywords in predicted queries. The improved version, DocT5Query (Cheriton, 2019), trains a T5 model to predict queries, but remains confined to sparse retrievers. There are also proposals to directly store question-answer pairs for retrieving and ranking, such as RePAQ (Lewis et al., 2021), QUADRo and QR (Campese et al., 2023, 2024). However, these attempts suffer from expanded index size, longer retrieval latency, and limited query handling. Hence, the query-centric alignment for dense retrievers remains challenging and largely unexplored.

3 Method

3.1 Problem Formulation

Given a query q and a document corpus $\mathcal{D} = \{d_1, d_2, ..., d_N\}$, our task is to retrieve a subset of K most relevant documents $\mathcal{D}_+ = \{d_{i_1}, d_{i_2}, ..., d_{i_K}\}$ through vector search. We define our embedding model as $\mathcal{E}(\cdot)$, which maps each document d and query q from the textual space $\mathcal{C}$ to a vector space $\mathbb{R}^r$. The semantic relevance is quantified by the cosine similarity, defined as: $sim(q,d) = \frac{\mathcal{E}(q)^T\mathcal{E}(d)}{\|\mathcal{E}(q)\|\|\mathcal{E}(d)\|}$. Furthermore, for each document d in our datasets, we invoke the query generator $\mathcal{Q}(\cdot)$ multiple times to generate n predicted queries $\{q_i\}_{i=1}^n$, where the cluster center in embedding space is captured by $\mathbb{E}[\mathcal{E}(\mathcal{Q}(d))]$ and Monte Carlo estimation $\overline{\mathcal{E}(\mathcal{Q}(d))}$.

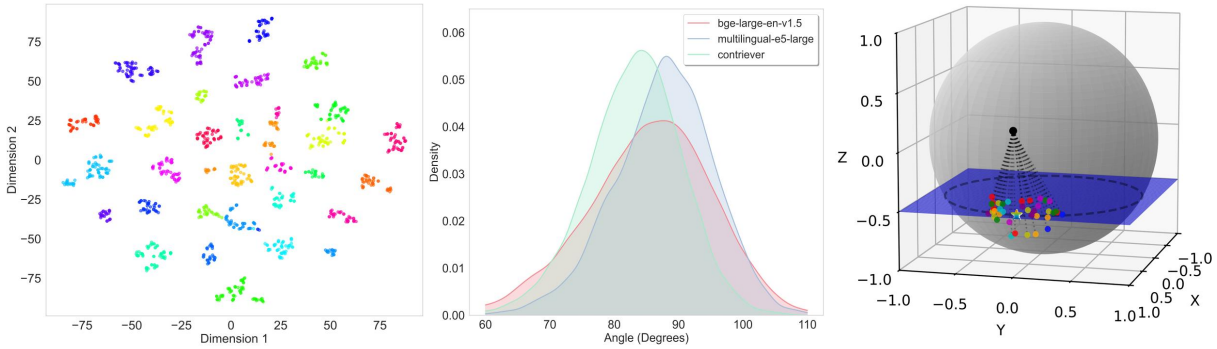
3.2 QAEncoder

We first introduce a novel encoding method, QAE_{base}, which represents the document by the cluster center $\mathbb{E}[\mathcal{E}(\mathcal{Q}(d))]$ of potential queries. Formally, the representation is defined as follows:

$$\begin{aligned}\text{QAE}_{\text{base}}(d) &= \mathbb{E}[\mathcal{E}(\mathcal{Q}(d))] \\ &\approx \overline{\mathcal{E}(\mathcal{Q}(d))} = \frac{1}{n}\sum_{i=1}^{n}\mathcal{E}(q_i).\end{aligned} \tag{1}$$

To thoroughly illustrate its mechanism, we formally define the *conical distribution hypothesis* and validate its reasonableness with empirical analysis.

Hypothesis 1 (Conical Distribution Hypothesis)**.** *For any document d, the potential queries approximately form a single cluster on some hyperplane $\mathcal{H} = \{x \in \mathbb{R}^r \mid w \cdot x = b\}$ in the semantic space, where $w \in \mathbb{R}^r$ is the normal vector and $b \in \mathbb{R}$ is*



(a) t-SNE visualization of the embedding space (b) Angular distribution of embeddings (c) Conical distribution in unit sphere

Figure 2: The conical distribution hypothesis. (a) t-SNE visualization of queries derived from various documents in the embedding space, illustrating distinct clustering behavior. (b) Angular distribution of document and query embeddings, showing the distribution of angles between $v_d = \mathcal{E}(d) - \mathbb{E}[\mathcal{E}(\mathcal{Q}(d))]$ and $v_{q_i} = \mathcal{E}(q_i) - \mathbb{E}[\mathcal{E}(\mathcal{Q}(d))]$. The angles form a bell curve just below 90°, supporting that v_d is approximately orthogonal to each v_{q_i}. (c) 3D visualization illustrating the conical distribution of the document (black point) and query (colored points) embeddings within a unit sphere. The star indicates the queries' cluster center.

the bias term. Furthermore, the document embedding $\mathcal{E}(d)$ lies on the perpendicular line intersecting the cluster center $\mathbb{E}[\mathcal{E}(\mathcal{Q}(d))]$. Formally, the relationship can be expressed as:

$$\mathcal{E}(d) \approx \mathbb{E}[\mathcal{E}(\mathcal{Q}(d))] + \lambda w, \quad \lambda \in \mathbb{R}.$$

We acknowledge that a more realistic model is an oblique cone, not the regular cone. Despite the hypothesis being highly simplified, it clearly captures the dynamics of QAEncoder's alignment. As shown in Fig. 2, the document embedding $\mathcal{E}(d)$ is an outlier of the cluster of potential queries; whereas $\text{QAE}_{\text{base}}(d)$ bridges this substantial gap, mathematically denoted by λw. We leave more discussions in Appendix A.2 for interested readers.

Nonetheless, the ideal embedding model should not only bring related entries closer but also separate unrelated entries as much as possible. Despite QAE_{base} increases document-query similarity, it poses the distinguishability issue.

3.2.1 Document Fingerprint Strategies

The distinguishability issue arises because incorporating too much query semantics into document representations suppresses their unique characteristics, and renders unrelated documents more similar. To address this issue, we reintroduce the characteristic of documents, intuitively termed *document fingerprints*, which enhance the uniqueness of QAE_{base} representations from different perspectives.

QAE_{base} + Embedding fingerprint = QAE_{emb}. The QAE_{emb} strategy manipulates within the embedding space and reintroduces unique identity of the original document, i.e. the document embedding $\mathcal{E}(d)$. Specifically, QAE_{emb} considers both the cluster center, QAE_{base}, and the document embedding, $\mathcal{E}(d)$, balancing their contributions using a hyperparameter α. The adjusted embedding is formulated as follows:

$$\begin{aligned}\text{QAE}_{\text{emb}}(d) &= (1-\alpha)\cdot\mathcal{E}(d) + \alpha\cdot\text{QAE}_{\text{base}}(d) \\ &\approx (1-\alpha)\cdot\mathcal{E}(d) + \alpha\cdot\frac{1}{n}\sum_{i=1}^{n}\mathcal{E}(q_i).\end{aligned} \tag{2}$$

QAE_{base} + Textual fingerprint = QAE_{txt}. The QAE_{txt} strategy focuses on the textual space and injects the document identity in a more straightforward manner. Let us define the length of text c as $|c|$. Before embedding, each document d is enriched by concatenating it with predicted queries to achieve a length ratio of approximately β between the queries and the original document. Then, the final embedding is derived as the average representation of these enriched documents.

$$\begin{aligned}d_i^* &= \text{concat}(d, \{q_j\}_{j=1}^{k}),\ \text{s.t. } |d_i^*| \approx (1+\beta)|d|. \\ \text{QAE}_{\text{txt}}(d) &= \frac{1}{n}\sum_{i=1}^{n}\mathcal{E}(d_i^*).\end{aligned} \tag{3}$$

Here k is dynamically determined by the constraint $|d_i| \approx (1+\beta)|d|$. For each enriched document d_i^*, we first shuffle predicted queries, and iteratively append them after d until reaching the length limit.

Hybrid fingerprint - QAE_{hyb}. The hybrid approach, QAE_{hyb}, seeks to combine the benefits of

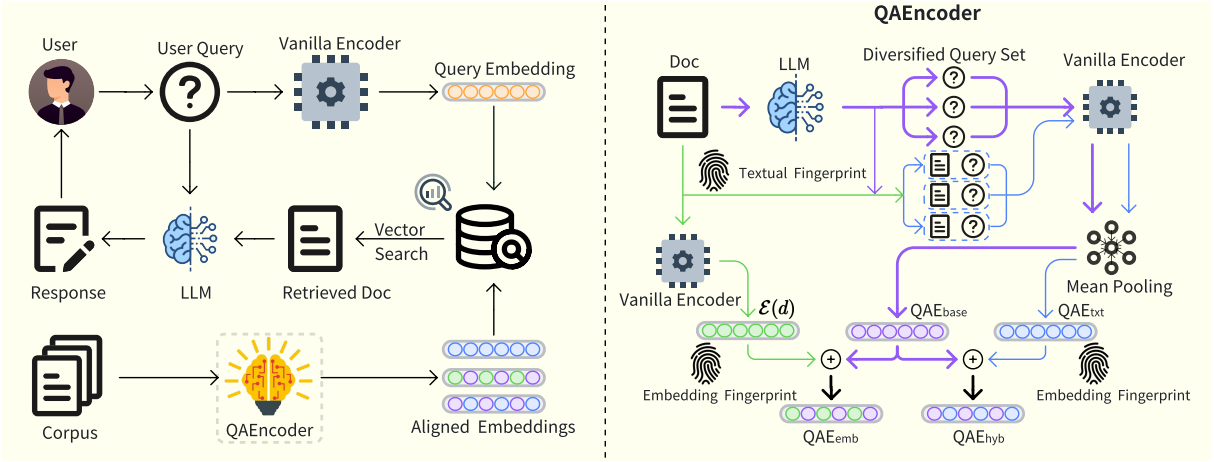


Figure 3: Architecture of QAEncoder. Left: Corpus documents are embedded using QAEncoder to obtain query-aligned representations for indexing. User queries are encoded with a vanilla encoder and used to retrieve relevant documents. Right: Internal mechanism of QAEncoder. QAEncoder addresses the document-query gap by generating a diverse set of queries for each document to create semantically aligned embeddings. Additionally, document fingerprint strategies are employed to ensure document distinguishability.

both QAE$_{emb}$ and QAE$_{txt}$. Although QAE$_{emb}$ integrates the document embedding $\mathcal{E}(d)$ and the cluster center $\mathbb{E}[\mathcal{E}(\mathcal{Q}(d))]$ through linear interpolation, inherent differences between these embeddings suggest that a simple linear interpolation should be suboptimal. Therefore, we explore the potential of substituting the document embedding $\mathcal{E}(d)$ in Equation 2 with QAE$_{txt}$, which fuses the semantics of both documents and queries.

$$\text{QAE}_{\text{hyb}}(d) = (1-\alpha)\cdot\text{QAE}_{\text{txt}}(d)+\alpha\cdot\text{QAE}_{\text{base}}(d). \quad (4)$$

In our implementation, all calculated embeddings are normalized for standardized cosine similarity search. Our experiments confirm QAE$_{emb}$ and QAE$_{hyb}$ outperform QAE$_{base}$ and QAE$_{txt}$.

4 Experiments

Datasets and Metrics. To rigorously assess the effectiveness of QAEncoder, we employ well-established BEIR benchmark (Thakur et al., 2021), which contains fifteen publicly available datasets to evaluate the general retrieval performance across diverse domains. However, classical datasets are frequently utilized for pre-training or fine-tuning embedding models [1], involving ArguAna, FEVER, FiQA, HotpotQA, MSMARCO, NQ, SciDocs and so on (all from BEIR). Hence, *classical datasets gradually fall short of objectively reflecting the generalized alignment capabilities for state-of-the-art models*, particularly in a rapidly evolving and updated knowledge base.

Recognizing this limitation, we further test on two latest news datasets, the Chinese dataset CRUD-RAG (Lyu et al., 2024) and the multilingual dataset FIGNEWS (Zaghouani et al., 2024) covering English, Arabic, French, Hindi and Hebrew. For BEIR benchmark, we report NDCG@10 as the evaluation metric like previous works (Thakur et al., 2021; Wang et al., 2021; Meng et al., 2022). For the latest datasets, we report both MRR@10 and NDCG@10 metrics, capturing both recall and ranking capabilities. We adopt the cheapest GPT-4o-mini as the main query generator, see Appendix B for query prediction pipeline and cost details.

Hyperparameter Setting. For the hyperparameters α and β, we define the following search spaces: $\alpha \in \{0.0, 0.15, 0.3, 0.45, 0.6, 0.75, 0.9\}$ and $\beta \in \{0.25, 0.5, 0.75, 1.0, 1.25, 1.5\}$. We adopt grid search for QAE$_{hyb}$. See Appendix H for more hyperparameter selection consideration.

For space reasons, we leave implementation details in Appendix D, the results on complete BEIR benchmark and FIGNEWS datasets in French, Hindi, and Hebrew in Appendix J.

4.1 Main Results

We mainly compare QAEncoder against the vanilla encoders, i.e. the backbones. Query-centric methods for sparse retrievers are also presented. Our comparison involves the following approaches:

Sparse retrievers - BM25 (Robertson and

[1]Please see the fine-tuning data of bge-m3 and bge-multilingual-gemma2-9b.

Model	Method	**AVG**	ArguAna	CQADups.	FEVER	MSMARCO	SciFact	Touche20	TRECC.
					Sparse				
BM25	-	46.9	31.5	29.9	**75.3**	22.8	66.5	**36.7**	65.6
DocT5Query	-	**49.4**	**34.9**	**32.5**	71.4	**33.8**	**67.5**	34.7	**71.3**
					Dense				
dpr	-	26.4	17.5	15.3	56.2	17.7	31.8	13.1	33.2
	QAE_{emb}, $\alpha = 0.45$	**37.1**	**29.6**	**22.3**	**70.9**	**25.4**	**40.4**	**22.9**	**48.0**
contriever	-	38.1	37.9	28.4	68.2	20.6	64.9	19.3	27.4
	QAE_{emb}, $\alpha = 0.45$	**45.8**	**47.1**	**33.5**	**73.1**	**26.1**	**70.2**	**25.2**	**45.2**
contriever-msmarco	-	49.0	44.6	34.5	75.8	**40.7**	67.7	20.4	59.6
	QAE_{hyb}, $\alpha = 0.3,\ \beta = 0.75$	**54.9**	**53.9**	**38.2**	**82.0**	39.9	**73.6**	**26.3**	**70.6**
bge-large-en-v1.5	-	58.5	63.5	42.2	87.2	**42.5**	74.6	24.8	74.8
	QAE_{hyb}, $\alpha = 0.15,\ \beta = 0.5$	**61.8**	**68.8**	**45.6**	**91.5**	41.2	**78.9**	**28.1**	**78.2**
multilingual-e5-large	-	55.0	54.4	39.7	**82.8**	**43.7**	70.4	23.1	71.2
	QAE_{hyb}, $\alpha = 0.15,\ \beta = 1.5$	**58.0**	**61.1**	**44.3**	82.1	43.0	**73.9**	**26.3**	**75.1**
e5-large-v2	-	52.9	46.4	37.9	82.8	**43.5**	72.2	20.7	66.5
	QAE_{hyb}, $\alpha = 0.3,\ \beta = 1.0$	**57.0**	**55.1**	**41.2**	**86.5**	42.8	**75.3**	**23.8**	**74.2**
gte-base-en-v1.5	-	59.4	63.5	39.5	**94.8**	**42.6**	76.8	25.2	73.1
	QAE_{hyb}, $\alpha = 0.3,\ \beta = 0.5$	**62.2**	**68.2**	**43.7**	94.2	41.9	**80.3**	**29.3**	**77.5**
jina-embeddings-v2-small-en	-	48.9	46.7	38.0	68.0	37.3	63.9	23.5	65.2
	QAE_{hyb}, $\alpha = 0.15,\ \beta = 0.5$	**54.2**	**55.3**	**42.3**	**74.3**	**39.5**	**67.2**	**27.2**	**73.6**

Table 1: Retrieval performance on seven BEIR benchmarks (NDCG@10). Hyperparameters including QAEncoder variants and weight terms α, β are optimized simultaneously for all datasets in BEIR. See Table 7 for the full table.

Zaragoza, 2009) and DocT5Query (Cheriton, 2019), an improved version of Doc2Query.

Dense retrievers - The state-of-the-art embedding models such as BGE models (Xiao et al., 2024) by BAAI, E5 models (Wang et al., 2023a) by Microsoft, GTE models (Zhang et al., 2024b) by Alibaba-NLP, Jina models (Günther et al., 2023) by Jina AI; Other well-known models like Contriever models (Izacard et al., 2021) by Facebook Research, BCEmbedding models (NetEase Youdao, 2023) by NetEase Youdao and the popular Text2Vec models (Xu, 2023). For more reference, we also include the seminal dense retriever DPR (Karpukhin et al., 2020), and models fine-tuned on Quora (Thakur et al., 2021), a large dataset of question pairs for question-query retrieval. We integrate them with QAEncoder to bridge the document-query gap.

More details can be found in Appendix F.

4.1.1 Performance on Classical Datasets

As shown in Table 1, for sparse retrievers, DocT5Query effectively augments documents with query prediction, and exhibits significant improvements on the BEIR benchmark over the standard BM25. Dense retrievers gradually outperform sparse retrievers with the scaling-up of model parameter and traning data. For the state-of-the-art embedding models such as BGE, E5, and Jina, integrating them with QAEncoder can lead to robust and generalized alignment, particularly for rare or unseen datasets. For instance, the jina-embeddings-v2-small-en model witnesses a NDCG increase from 65.2 to 73.6 on TRECC dataset; the bge-large-en-v1.5 model's NDCG on SciFact dataset rises from 74.6 to 78.9; the e5-large-v2 model achieves 8.7 NDCG gains on the ArguAna dataset, as the ArguAna dataset is not included in the fine-tuning data (Wang et al., 2022).

Moreover, QAEncoder significantly improves the performance of other well-known models. For example, the contriever model and its fine-tuned version, contriever-msmarco, achieve 17.8 and 11 NDCG gains on TRECC dataset respectively. More data is available in Tab. 7.

4.1.2 Performance on Latest Datasets

In scenarios such as search engine, financial analysis, and news QA, large volumes of new data constantly emerge and are indexed into retrieval base for accurate and up-to-date response. Hence, the alignment capability for previously unseen user queries and relevant documents is crucial for embedding models in RAG systems. We experiment on the latest news datasets, FIGNEWS and CRUD-RAG, to avoid data leakage and mimic the real-world scenarios. As illustrated in Table 2, for latest datasets, QAEncoder significantly improves across both state-of-the-art embedding models and other well-known models. E.g., the gte-multilingual-base

Model	Method	FIGNEWS(English)		FIGNEWS(Arabic)		CRUD-RAG(Chinese)	
		MRR@10	NDCG@10	MRR@10	NDCG@10	MRR@10	NDCG@10
bge-m3	-	74.4	78.7	77.8	80.9	47.5	48.6
	QAE_{txt}, $\beta = 1.5$	**77.2**	**81**	**80.2**	**83.1**	**51.4**	**52.5**
multilingual-e5-small	-	71	75.1	74.1	77.4	44.6	46.0
	QAE_{hyb}, $\alpha = 0.3,\ \beta = 0.5$	**74.6**	**78.5**	**78.9**	**81.6**	**50.6**	**51.6**
multilingual-e5-base	-	74.8	78.1	72.3	76	47.0	48.2
	QAE_{emb}, $\alpha = 0.3$	**77.6**	**81.3**	**77.2**	**80.3**	**51.2**	**52.3**
multilingual-e5-large	-	73.9	77.8	76.7	80.2	46.9	48.3
	QAE_{hyb}, $\alpha = 0.15,\ \beta = 1.25$	**77.1**	**80.6**	**82.2**	**85.1**	**51.5**	**52.7**
gte-multilingual-base	-	65.5	70.4	73.4	76.8	45.3	46.8
	QAE_{hyb}, $\alpha = 0.15,\ \beta = 1.5$	**75.5**	**79.5**	**76.2**	**79.1**	**49.4**	**51.0**
mcontriever	-	32.9	36.7	40.3	44.7	39.2	41.6
	QAE_{hyb}, $\alpha = 0.45,\ \beta = 1.25$	**61.4**	**65.9**	**68.3**	**72.1**	**51.3**	**52.4**
bce-embedding-base-v1	-	59.1	63.8	-	-	42.0	44.0
	QAE_{hyb}, $\alpha = 0.3,\ \beta = 0.5$	**66.8**	**71.1**	-	-	**49.7**	**51.0**
text2vec-base-multilingual	-	38.7	43.6	27.8	31.9	9.7	10.6
	QAE_{emb}, $\alpha = 0.75$	**55.4**	**59.9**	**51.5**	**55.4**	**32.1**	**34.1**
quora-distilbert-multilingual	-	39.8	44.2	28.2	32.1	21.8	23.5
	QAE_{emb}, $\alpha = 0.6$	**50.5**	**54.7**	**38.7**	**42.8**	**34.4**	**36.8**

Table 2: Retrieval performance on the latest datasets FIGNEWS and CRUD-RAG. Hyperparameters including QAEncoder variants and weight terms α, β are optimized simultaneously for all lastest datasets. We leave the full table in Tab. 8, and results on monolingual and bilingual embedding models in Tab. 11 for interested readers.

model and the mcontriever model's MRR metrics increase from 65.5 to 75.5, and 32.9 to 61.4 respectively on FIGNEWS(English) dataset. The multilingual-e5-large model's MRR increase from 76.7 to 82.2 on FIGNEWS(Arabic). Additionally, the text2vec-base-multilingual model's MRR on CRUD-RAG dataset rises from 9.7 to 32.1, while the bge-m3 model's MRR improves from 47.5 to 51.4. These results remarkably confirm QAEncoder's generalized alignment capability across various embedding models and languages. See Tables 8 and 11 for more embedding models and datasets.

4.2 Analysis and Discussion

For more comprehensive assessments, we analyze various QAEncoder ablations, i.e. QAE_{emb}, QAE_{txt}, QAE_{hyb} and $\text{QAE}_{\text{naive}}$, which directly stores predicted queries. We also evaluate QAEncoder's robustness with respect to the query generator. Finally, we discuss the relationship between QAEncoder and both training-based and document-centric methods.

4.2.1 Ablations of QAEncoder

We present the performance comparison of QAEncoder variants on two state-of-the-art embedding models in Table 3. Generally, QAE_{hyb} and QAE_{emb} outperform the QAE_{txt} and $\text{QAE}_{\text{naive}}$ approaches. For instance, for the bge-m3 model, QAE_{hyb} consistently outperforms other variants. Conversely, the multilingual-e5-large model performs best with QAE_{emb}. However, the best performance differences between QAE_{emb}, QAE_{txt}, and QAE_{hyb} are not substantial, demonstrating the robustness of our approach to hyperparameter variations. Regarding $\text{QAE}_{\text{naive}}$, it evidently underperforms other ablations, despite storing 10 times the number of embedding vectors. This leads to unacceptable storage management overhead and recall latency in large-scale production systems. We provide more granular ablation experiments in Fig. 4, as well as the convergence speed of QAE_{base}'s Monte Carlo estimation. See Table 9 for the full table.

4.2.2 Robustness w.r.t. Query Generator

The calculation of $\text{QAE}_{\text{base}} \approx \frac{1}{n}\sum_{i=1}^{n}\mathcal{E}(q_i)$ is largely robust to query generators, thanks to the nature of Monte Carlo estimation (Hsu and Robbins, 1947). As shown in Fig. 5, for any given document and embedding model (e.g. BGE and E5), the QAE_{base} representations derived from different query generators are quite consistent. For example, the worst cosine similarity remains as high as 0.95 for the bge-large-en-v1.5 encoder. Note that QAE_{emb}, QAE_{txt}, and QAE_{hyb} integrate document fingerprints into QAE_{base}, and thus have better query generator independence.

Model	Method	FIGNEWS(English)		FIGNEWS(Arabic)		CRUD-RAG(Chinese)	
		MRR@10	NDCG@10	MRR@10	NDCG@10	MRR@10	NDCG@10
bge-m3	QAE_{emb}, $\alpha = 0.3$	76.4	80.5	80.1	82.9	51.3	52.4
	QAE_{txt}, $\beta = 1.5$	77.2	81	80.2	83.1	51.4	52.5
	QAE_{hyb}, $\alpha = 0.15,\ \beta = 1.5$	**77.4**	**81.1**	**80.6**	**83.4**	**51.7**	**52.7**
	$\text{QAE}_{\text{naive}}$, n=10	76.9	79.9	77.1	79.7	47.0	48.1
multilingual-e5-large	QAE_{emb}, $\alpha = 0.45$	**77.9**	**81.4**	79.8	83	**51.9**	**52.9**
	QAE_{txt}, $\beta = 1.5$	75.6	79.2	80.9	84.1	51.0	52.3
	QAE_{hyb}, $\alpha = 0.15,\ \beta = 1.25$	77.1	80.6	**82.2**	**85.1**	51.5	52.7
	$\text{QAE}_{\text{naive}}$, n=10	77.5	80.3	76.5	79.4	46.5	47.7

Table 3: Performance comparison of QAEncoder variants on the latest datasets. Hyperparameters are optimized simultaneously for all the latest datasets. n indicates the number of predicted queries in $\text{QAE}_{\text{naive}}$.

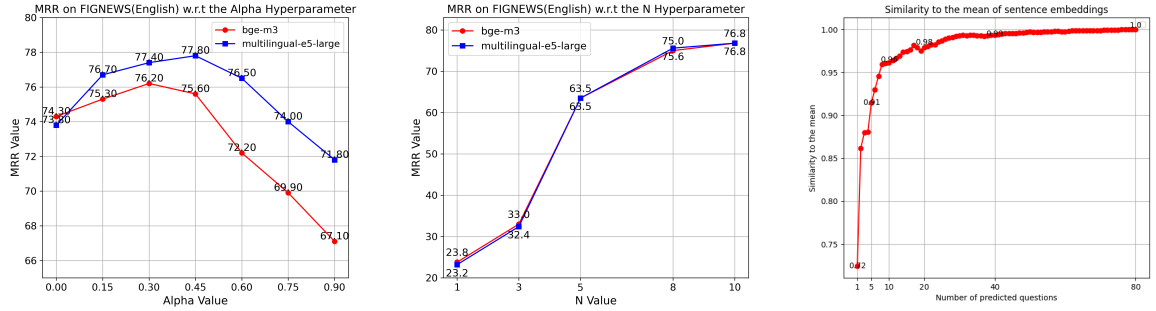


Figure 4: Left: The impact of varying α values for QAE_{emb}. Middle: The effect of varying the number of predicted queries for $\text{QAE}_{\text{naive}}$, with MRR improving as n increases, approaching stability at $n = 10$. The curves for different models are mostly identical as the matching are largely driven by predicted queries. Right: The convergence speed for Monte Carlo estimation of QAE_{base}. Please refer to Appendix C and Fig. 11 for more cost details.

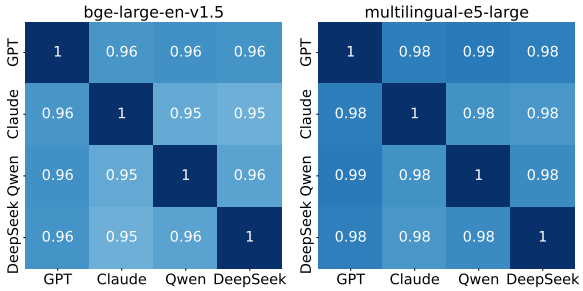


Figure 5: Similarity matrices of QAE_{base} with different query generators (average value on 100 documents and 10 queries per document), exhibiting high consistency.

Model	Method	GPT	Claude	Qwen	DeepSeek
bge-large-en-v1.5	QAE_{emb}, $\alpha = 0.5$	72.2	71.6	71.0	70.5
	QAE_{txt}, $\beta = 1.5$	72.7	73.5	72.6	72.5
	QAE_{hyb}, $\alpha = 0.15, \beta = 1.5$	73.7	73.5	73.2	73.3
multilingual-e5-large	QAE_{emb}, $\alpha = 0.5$	77.6	76.1	75.8	75.9
	QAE_{txt}, $\beta = 1.5$	75.1	74.9	74.7	74.6
	QAE_{hyb}, $\alpha = 0.15, \beta = 1.5$	77.0	76.5	76.6	76.6

Table 4: MRR@10 on FIGNEWS(English) with GPT-generated test queries and different query predictors. The performance is largely robust to query predictors.

As shown in Tab. 4, we adopted GPT-4o-mini to generate test queries and various LLMs as generators for query prediction. For QAE_{emb}, GPT-4o-mini shows a slight advantage over other predictors. But for QAE_{txt} and QAE_{hyb}, the performance is largely consistent, highlighting not only the importance of document fingerprints but also the robustness of QAEncoder. Interestingly, Claude beats GPT on QAE_{txt} with MRR values 73.5 and 72.7 respectively, while the vanilla BGE is 66.1. The full names of abbreviations are in left Appendix B.

4.2.3 Training-based and Document-centric Methods

Training-based approaches mainly include two types: fine-tuning on QA datasets (domain adaptation) and fine-tuning on multi-task instruction datasets. We select mcontriever-msmarco and multilingual-e5-large-instruct as representative methods respectively. To unveil the catastrophic forgetting issue of training-based methods, we incorporate GPL (Wang et al., 2021), which predicts queries, mines hard negative samples, and distills the re-ranker for unsupervised domain adaptation.

For multi-domain adaptation, we fine-tune on FIGNEWS and CRUD-RAG datasets iteratively to simulate knowledge update. While GPL bridges

Model	Method	FIGNEWS(English)		FIGNEWS(Arabic)		CRUD-RAG(Chinese)	
		MRR@10	NDCG@10	MRR@10	NDCG@10	MRR@10	NDCG@10
mcontriever	-	32.9	36.7	40.3	44.7	39.2	41.6
	QAE_{hyb}, $\alpha = 0.45,\ \beta = 1.25$	61.4	65.9	68.3	72.1	51.3	52.4
	GPL†	68.3	73.6	72.1	75.18	**52.8**	**55.9**
	MS†	66.1	70.6	70.2	73.7	46.5	47.8
	MS† + QAE_{hyb}, $\alpha = 0.3,\ \beta = 0.75$	**72.3**	**76.8**	**77.3**	**80.5**	51.2	52.4
	QA-RAG‡	31.1	34.4	42.3	46.8	43.8	45.8
	Query2Doc‡	25.1	29.0	34.5	38.9	35.7	37.2
	HyDE‡	25	27.9	35.7	41.9	36.7	38.7
multilingual-e5-large	-	73.9	77.8	76.7	80.2	46.9	48.3
	QAE_{hyb}, $\alpha = 0.15,\ \beta = 1.25$	**77.1**	**80.6**	**82.2**	**85.1**	51.5	52.7
	GPL†	75.2	78.9	79.4	82.3	**53.6**	**56.3**
	INS†	67	71.4	75	78.2	43.7	45.2
	INS† + QAE_{hyb}, $\alpha = 0.15,\ \beta = 1.5$	75.6	79.8	80.8	83.7	51.4	52.4
	QA-RAG‡	73.3	76.5	72.8	75.6	45.8	46.5
	Query2Doc‡	63.4	68.2	66.3	72.8	42.0	43.1
	HyDE‡	63.6	68.3	68.3	74.1	42.3	43.6

Table 5: Performance comparison of QAEncoder with training-based and document-centric methods on the latest datasets FIGNEWS and CRUD-RAG. Hyperparameters are optimized simultaneously for all latest datasets. † indicates the training-based methods. MS† represents fine-tuning on MSMARCO, i.e. the mcontriever-msmarco model; INS† represents instruction-tuning, i.e. the multilingual-e5-large-instruct model (Wang et al., 2023a). ‡ indicates the document-centric methods. See Table 10 for the full table.

the document-query gap with notable improvements on CRUD-RAG dataset, the catastrophic forgetting issue is serious: GPL improves marginally on FIGNEWS datasets, while QAEncoder works robustly due to the training-free nature.

Training-based and query-centric methods operate at training time and indexing time, respectively. Therefore, integrating these approaches could lead to more improvements. As illustrated in Table 5, both types of fine-tuned models significantly benefit from the QAEncoder. For instance, the mcontriever-msmarco model improves MRR from 70.2 to 77.3 on FIGNEWS(Arabic); the multilingual-e5-large-instruct model's MRR increases 8.6 and 7.7 MRR points on the FIGNEWS(English) and CRUD-RAG(Chinese) datasets, respectively.

For document-centric methods, we consider HyDE (Gao et al., 2023), Query2Doc (Wang et al., 2023b), and QA-RAG (Kim and Min, 2024) for comparison. Note that the pseudo-document generator in QA-RAG is fine-tuned on medical QA datasets, here we adopt out-of-box GPT-4o-mini. The widely-reported hallucination phenomenon on the latest datasets is confirmed (Wang et al., 2023b). As shown in Tab. 5, the retrieval performance of HyDE and Query2Doc heavily decreases for all the latest datasets, attributed to the hallucination of pseudo-document generation. Although QA-RAG alleviates hallucination by its two-way retrieval and re-ranking mechanism, it still underperforms. Besides, the LLM invocation for pseudo-documents is both costly and time-consuming. In our case, the time for single LLM invocation is more than 2000ms while the time for vector search is less than 10ms. These highlight the irreplaceable importance and practicality of QAEncoder method.

5 Conclusion

In this paper, we propose QAEncoder to bridge the document-query gap from the query-centric perspective—a novel, training-free and pioneering approach. QAEncoder replaces document embeddings with the expectation of query embeddings, theoretically supported by the conical distribution hypothesis and practically enhanced by document fingerprint strategies. Extensive experiments are conducted on both classical BEIR benchmark suite and the latest news datasets, covering 20+ embedding models and 6 languages. QAEncoder demonstrates robust generalization, diverse query handling, and compatibility with existing RAG architectures and training-based methods.

Acknowledgments

This work is supported by the National Key R&D Program of China (2024YFA1014003), National Natural Science Foundation of China (92470121, 62402016), CAAI-Ant Group Research Fund, and High-performance Computing Platform of Peking University.

Limitations

Despite the benefits of QAEncoder, there are still ongoing works for further improvements:

- The current mean pooling could be overly simplistic and limits the performance improvement. The multi-cluster version such as Gaussian mixture models and multi-vector representation could be explored.

- Since the query generator mainly generates simple queries, out-of-domain issues with complex, multi-hop queries could happen. However, query optimization has become increasingly prevalent to refine the original user query, making it more suitable for retrieval module. Practically, query decomposition and rewriting can be integrated to break down complex queries into simpler sub-queries and to rewrite these sub-queries into the in-domain style, respectively. Therefore, we focus QAEncoder on the core retrieval module, and leave the handling of complex queries to auxiliary strategies.

- The 5W1H framework captures our core design principles and serves as the motivation for our work, primarily applicable to narrative or factual texts. Without loss of generality, QAEncoder estimates the cluster center of potential queries via Monte Carlo method, where the type of query is arbitrary and not constrained to 5W1H. We acknowledge there could be the worst case, user query may deviate from predicted ones, i.e. the outlier. Therefore, more fallback mechanisms beyond document fingerprints possess research value.

- There is also a risk of data leakage for property information when predicting queries via API calls. The query prediction process may incur underlying security vulnerabilities.

We aim to keep QAEncoder simple-yet-effective, and leave these problems in the future research.

References

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

Paulo Alting von Geusau and Peter Bloem. 2020. Evaluating the robustness of question-answering models to paraphrased questions. In *Benelux Conference on Artificial Intelligence*, pages 1–14. Springer.

Akari Asai, Zeqiu Wu, Yizhong Wang, et al. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arxiv:2310.11511*.

Alexander Bondarenko, Maik Fröbe, Johannes Kiesel, Shahbaz Syed, Timon Gurcke, Meriem Beloucif, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, et al. 2022. Overview of touché 2022: argument retrieval. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 311–336. Springer.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38*, pages 716–722. Springer.

Stefano Campese, Ivano Lauriola, and Alessandro Moschitti. 2023. Quadro: Dataset and models for question-answer database retrieval. *arXiv preprint arXiv:2304.01003*.

Stefano Campese, Ivano Lauriola, and Alessandro Moschitti. 2024. Pre-training methods for question reranking. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 469–476.

David R. Cheriton. 2019. From doc2query to doctttttquery.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.

Zhe Dong, Jianmo Ni, Daniel M Bikel, Enrique Alfonseca, Yuan Wang, Chen Qu, and Imed Zitouni. 2022. Exploring dual encoder architectures for question answering. In *EMNLP*, pages 9414–9419.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Thibault Févry, Livio Baldini Soares, et al. 2020. Entities as experts: Sparse memory access with entity supervision. In *EMNLP*.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *ACL*.

Michael Günther, Jackmin Ong, Isabelle Mohr, Alaeddine Abdessalem, Tanguy Abel, Mohammad Kalim Akram, Susana Guzman, Georgios Mastrapas, Saba Sturua, Bo Wang, et al. 2023. Jina embeddings 2: 8192-token general-purpose text embeddings for long documents. *arXiv preprint arXiv:2310.19923*.

Quan Guo, Shuai Cao, and Zhang Yi. 2022. A medical question answering system using large language models and knowledge graphs. *International Journal of Intelligent Systems*, 37(11).

Aman Gupta, Anup Shirgaonkar, Angels de Luis Balaguer, Bruno Silva, Daniel Holstein, Dawei Li, Jennifer Marsman, Leonardo O Nunes, Mahsa Rouzbahman, Morris Sharp, et al. 2024. Rag vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture. *arXiv preprint arXiv:2401.08406*.

Kelvin Guu, Kenton Lee, Zora Tung, et al. 2020. REALM: retrieval-augmented language model pretraining. *ICML*.

Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. 2017. Dbpedia-entity v2: a test collection for entity search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1265–1268.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 113–122.

Doris Hoogeveen, Karin M Verspoor, and Timothy Baldwin. 2015. Cqadupstack: A benchmark data set for community question-answering research. In *Proceedings of the 20th Australasian document computing symposium*, pages 1–8.

Pao-Lu Hsu and Herbert Robbins. 1947. Complete convergence and the law of large numbers. *Proceedings of the national academy of sciences*, 33(2):25–31.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Hiroki Iida and Naoaki Okazaki. 2022. Unsupervised domain adaptation for sparse retrieval by filling vocabulary and word frequency gaps. *arXiv preprint arXiv:2211.03988*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *EACL*.

Tony Jinks. 2019. *The 5W1H Method*, page 41–44. Springer International Publishing.

Karen Sparck Jones. 1973. Index term weighting. *Information storage and retrieval*, 9(11):619–633.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Sujit Khanna and Shishir Subedi. 2024. Tabular embedding model (tem): Finetuning embedding models for tabular rag applications. *arXiv preprint arXiv:2405.01585*.

Jaewoong Kim and Moohong Min. 2024. From rag to qa-rag: Integrating generative ai for pharmaceutical regulatory compliance process. *arXiv preprint arXiv:2402.01717*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Zachary Levonian, Chenglu Li, Wangda Zhu, Anoushka Gade, Owen Henkel, Millie-Ellen Postle, and Wanli Xing. 2023. Retrieval-augmented generation to improve math question-answering: Trade-offs between groundedness and human preference. *arXiv preprint arXiv:2310.03184*.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. Paq: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*.

Zehan Li, Nan Yang, Liang Wang, and Furu Wei. 2022. Learning diverse document representations with deep query interactions for dense retrieval. *arXiv preprint arXiv:2208.04232*.

Yuanjie Lyu, Zhiyu Li, Simin Niu, et al. 2024. CRUD-RAG: A comprehensive chinese benchmark for retrieval-augmented generation of large language models. *arxiv:2401.17043*.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pages 1941–1942.

Subhankar Maity and Aniket Deroy. 2024. The future of learning in the age of generative ai: Automated question generation and assessment with large language models. *arXiv preprint arXiv:2410.09576*.

Antonio Mallia, Omar Khattab, Torsten Suel, and Nicola Tonellotto. 2021. Learning passage impacts for inverted indexes. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1723–1727.

Kanti V Mardia, John T Kent, and Charles C Taylor. 2024. *Multivariate analysis*, volume 88. John Wiley & Sons.

Goeffrey J McLachlan. 1999. Mahalanobis distance. *Resonance*, 4(6):20–26.

Rui Meng, Ye Liu, Semih Yavuz, Divyansh Agarwal, Lifu Tu, Ning Yu, Jianguo Zhang, Meghana Bhat, and Yingbo Zhou. 2022. Augtriever: Unsupervised dense retrieval by scalable data augmentation. *arXiv preprint arXiv:2212.08841*.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR*.

Inc. NetEase Youdao. 2023. Bcembedding: Bilingual and crosslingual embedding for rag. https://github.com/netease-youdao/BCEmbedding.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*.

Bolanle Ojokoh and Emmanuel Adebisi. 2018. A review of question answering systems. *Journal of Web Engineering*, 17(8):717–758.

James Jie Pan, Jianguo Wang, and Guoliang Li. 2024. Survey of vector database management systems. *The VLDB Journal*, pages 1–25.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. In *NAACL-HLT*. Association for Computational Linguistics.

Qwen Development Team. 2024a. Qwen 2.5 Speed Benchmark. https://qwen.readthedocs.io/en/latest/benchmark/speed_benchmark.html. Accessed: 2024-12-16.

Qwen Development Team. 2024b. Qwen 2.5 Turbo - Advanced Features and Improvements. https://qwen2.org/qwen2-5-turbo/. Accessed: 2024-12-16.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*.

Englevert P Reyes, Ron Mhel Francis L Blanco, Defanee Rose L Doroon, Jay Lord B Limana, and Ana Marie A Torcende. 2021. Feynman technique as a heutagogical learning strategy for independent and remote learning. *Recoletos Multidisciplinary Research Journal*, 9(2):1–13.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *FTIR*, 3(4):333–389.

Danielle Saunders. 2022. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *Journal of Artificial Intelligence Research*, 75:351–424.

Shamane Siriwardhana, Rivindu Weerasekera, Tharindu Kaluarachchi, et al. 2023. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *TACL*, 11:1–17.

Abhijit Suprem and Calton Pu. 2022. Evaluating generalizability of fine-tuned models for fake news detection. *arXiv preprint arXiv:2205.07154*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Yung Liang Tong. 2012. *The multivariate normal distribution*. Springer Science & Business Media.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. Trec-covid: constructing a pandemic information retrieval test collection. In *ACM SIGIR Forum*, volume 54, pages 1–12. ACM New York, NY, USA.

Xing W, Guangyuan Ma, Wanhui Qian, Zijia Lin, and Songlin Hu. 2023. Query-as-context pre-training for dense passage retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1906–1916, Singapore. Association for Computational Linguistics.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*.

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2021. Gpl: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. *arXiv preprint arXiv:2112.07577*.

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

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023a. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

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

Liang Wang, Nan Yang, and Furu Wei. 2023b. Query2doc: Query expansion with large language models. In *EMNLP*.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

Ming Xu. 2023. Text2vec: Text to vector toolkit. `https://github.com/shibing624/text2vec`.

Zhentao Xu, Mark Jerome Cruz, Matthew Guevara, Tie Wang, Manasi Deshpande, Xiaofeng Wang, and Zheng Li. 2024. Retrieval-augmented generation with knowledge graphs for customer service question answering. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2905–2909.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Wajdi Zaghouani, Mustafa Jarrar, Nizar Habash, Houda Bouamor, Imed Zitouni, Mona Diab, Samhaa R El-Beltagy, and Muhammed AbuOdeh. 2024. The fignews shared task on news media narratives. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 530–547.

Mingtian Zhang, Shawn Lan, Peter Hayes, and David Barber. 2024a. Mafin: Enhancing black-box embeddings with model augmented fine-tuning. *arXiv preprint arXiv:2402.12177*.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024b. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *Preprint*, arXiv:2407.19669.

Yueheng Zhang, Xiaoyuan Liu, Yiyou Sun, Atheer Alharbi, Hend Alzahrani, Basel Alomair, and Dawn Song. 2025. Can llms design good questions based on context? *arXiv preprint arXiv:2501.03491*.

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

Xi Zhao, Yao Tian, Kai Huang, Bolong Zheng, and Xiaofang Zhou. 2023. Towards efficient index construction and approximate nearest neighbor search in high-dimensional spaces. *Proceedings of the VLDB Endowment*, 16(8):1979–1991.

Zhi Zheng, Kai Hui, Ben He, Xianpei Han, Le Sun, and Andrew Yates. 2020. Bert-qe: Contextualized query expansion for document re-ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4718–4728.

Appendix

A Conical Distribution Hypothesis

A.1 Case Studies

Here we provide several concrete case studies in Fig. 6, 7, 8, 9. The conclusions are highly consistent with experiments in the main text, across embedding models and languages. Please refer to our repository for related scripts and reproduction.

A.2 Proof of Conical Distribution Hypothesis

This subsection provides the proof of the Conical Distribution Hypothesis, which proposes that potential queries form a distinct cluster on a hyperplane in semantic space.

Proof. Our validation is structured from three core aspects:

- **Single-cluster sub-hypothesis verification.** As illustrated in Fig. 2(a), we validate the single-cluster sub-hypothesis by visualizing the embedding space using t-SNE dimensionality reduction techniques. This visualization displays that the predicted queries for each document form distinct and cohesive clusters (different colored). And these clusters are notably distant from the clusters of other documents, thereby supporting the single-cluster sub-hypothesis.

- **Perpendicular sub-hypothesis verification.** To further assess the perpendicular sub-hypothesis, let $v_d = \mathcal{E}(d) - \mathbb{E}[\mathcal{E}(\mathcal{Q}(d))]$ and $v_{q_i} = \mathcal{E}(q_i) - \mathbb{E}[\mathcal{E}(\mathcal{Q}(d))]$ be the vectors from the cluster center to the document embedding and the individual query embedding, respectively. As illustrated in Fig. 2(b), the degree distribution between vector v_d and vector v_{q_i} exhibits a bell-shaped curve. The mean value is slightly less than 90 degrees, and the primary range of distribution lies between 75 and 100 degrees, which confirms that v_d is approximately orthogonal to each v_{q_i} and can be regarded as the normal vector to some hyperplane $\mathcal{H}$.

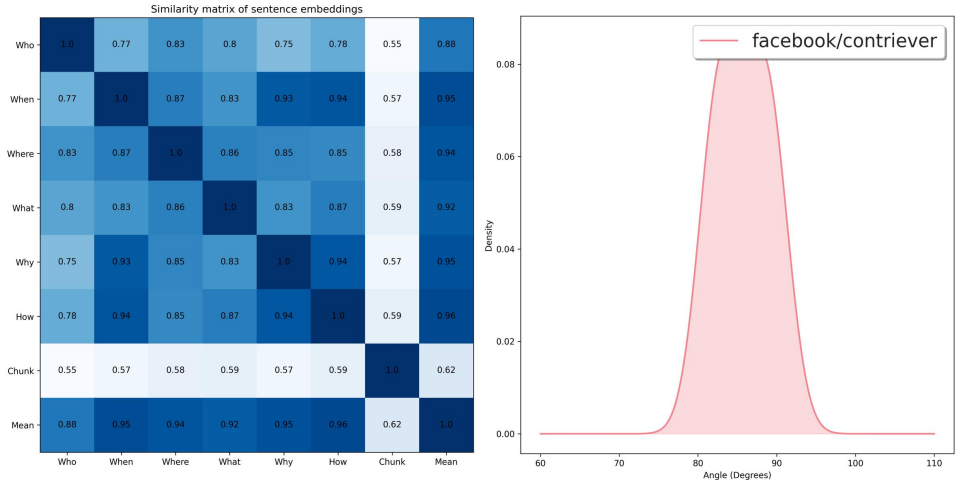
- **Conical distribution in unit sphere demonstration.** Finally, we illustrate the highly simplified conical distribution hypothesis within the unit sphere embedding space, as most embedding models utilize normalized embedding vectors. As depicted in Fig. 2(c), the embeddings of potential queries form a cluster on the surface of the unit sphere, with each point color-coded. The center of the cluster is indicated by a star, while the document embedding is represented by a black point positioned above the cluster. It is evident that these elements form a distorted cone, aligning with the above hypothesis and the degree distribution experiment.

□

Furthermore, when the stronger hypothesis assuming the cluster follows the Gaussian distribution is adopted, more quantitative analysis results can be derived.

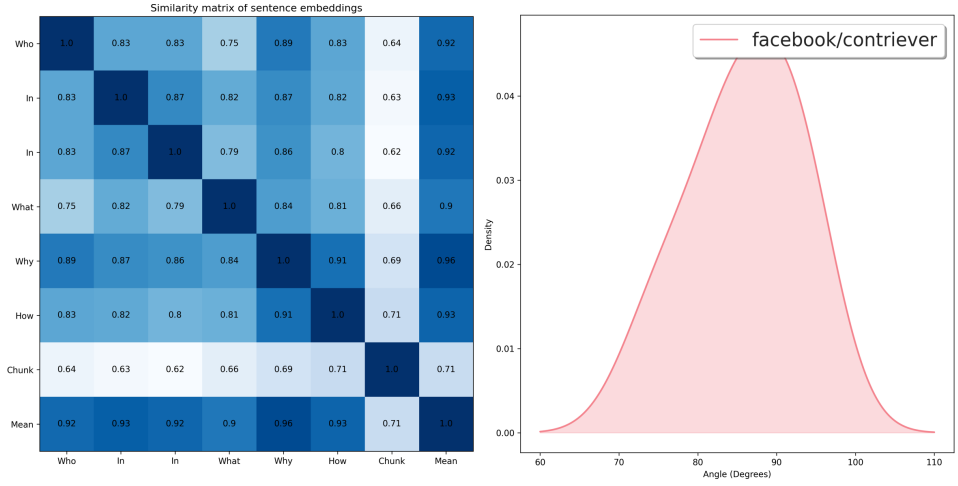
A.3 Strong Conical Distribution Hypothesis

In this subsection, we further substantiate the original hypothesis that the potential queries adhere to



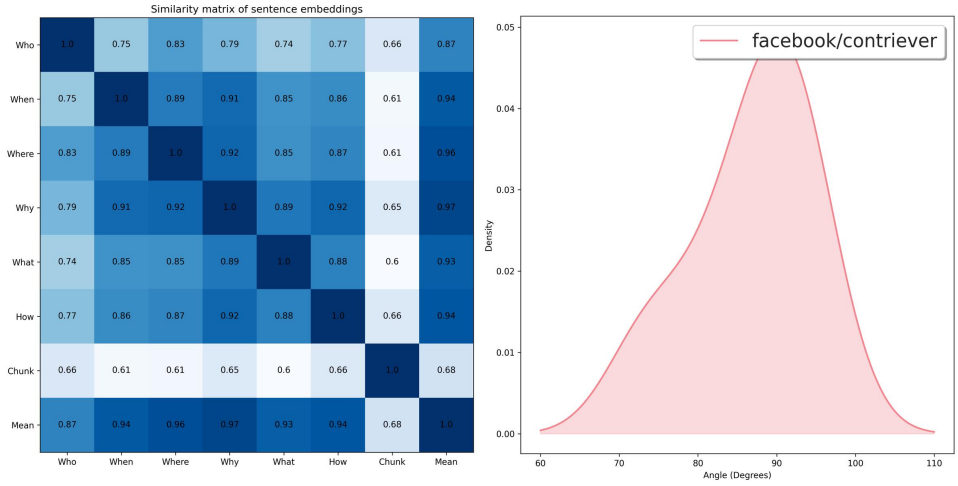
"Who was the astronaut that landed on the moon as part of the Apollo 11 mission?",
"When did Neil Armstrong land on the moon?",
"Where did Neil Armstrong land as part of the Apollo 11 mission?",
"What did Neil Armstrong achieve during the Apollo 11 mission?",
"Why did Neil Armstrong land on the moon?",
"How did Neil Armstrong successfully land on the moon?",
"On July 20, 1969, Neil Armstrong landed on the moon to accomplish the Apollo 11 mission, which was achieved through NASA's extensive planning and the Saturn V rocket."

Figure 6: Similarity matrix and angle distribution on **Armstrong example with contriever encoder**.



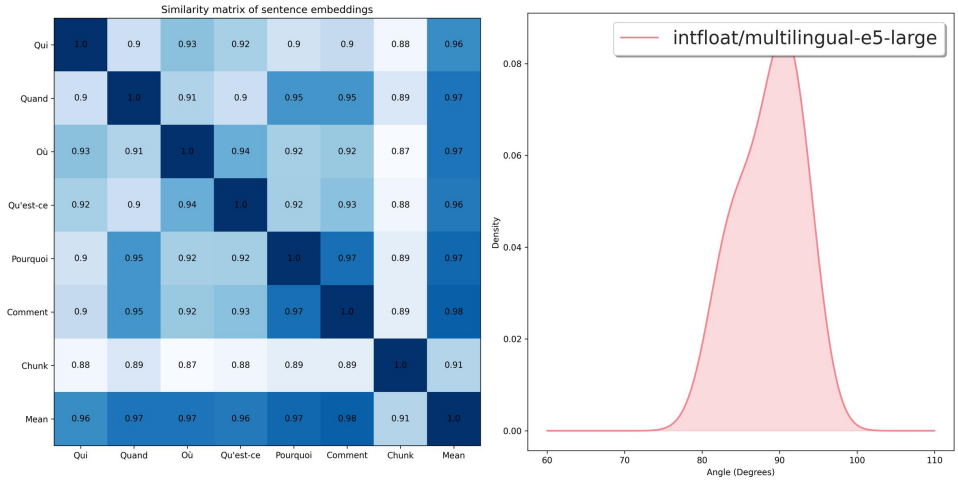
"Who developed the theory of general relativity?",
"In what year did Albert Einstein develop the theory of general relativity?",
"In which city did Albert Einstein develop the theory of general relativity?",
"What scientific theory did Albert Einstein develop in 1915?",
"Why did Albert Einstein develop the theory of general relativity?",
"How did Albert Einstein explain the force of gravity in his theory of general relativity?",
"Albert Einstein, in 1915 in Berlin, developed the theory of general relativity to explain the force of gravity by describing the curvature of spacetime caused by mass and energy."

Figure 7: Similarity matrix and angle distribution on **Einstein example with contriever encoder**.



"Who won the Nobel Prize in Physics in 1903 for their research on radioactivity?",
"When did Marie Curie win the Nobel Prize in Physics?",
"Where was the Nobel Prize in Physics awarded to Marie Curie in 1903?",
"Why was Marie Curie awarded the Nobel Prize in Physics in 1903?",
"What prestigious award did Marie Curie receive in 1903?",
"How did Marie Curie conduct her research that led to winning the Nobel Prize in Physics in 1903?",
"Marie Curie won the Nobel Prize in Physics in 1903 in Stockholm for her research on radioactivity by conducting experiments with uranium and radium."

Figure 8: Similarity matrix and angle distribution on **Curie example with contriever encoder**.



"Qui était l'astronaute qui a atterri sur la lune dans le cadre de la mission Apollo 11 ?",
"Quand Neil Armstrong a-t-il atterri sur la lune ?",
"Où Neil Armstrong a-t-il atterri dans le cadre de la mission Apollo 11 ?",
"Qu'est-ce que Neil Armstrong a accompli lors de la mission Apollo 11 ?",
"Pourquoi Neil Armstrong a-t-il atterri sur la lune ?",
"Comment Neil Armstrong a-t-il réussi à atterrir sur la lune ?",
"Le 20 juillet 1969, Neil Armstrong a atterri sur la lune pour accomplir la mission Apollo 11, qui a été réalisée grâce à la planification approfondie de la NASA et à la fusée Saturn V."

Figure 9: Similarity matrix and angle distribution on **Armstrong example in French with multilingual-e5-large**.

a Gaussian distribution: For any document d, the potential queries in the embedding space approximately follow a Gaussian distribution, characterized by a mean μ and covariance matrix Σ. Refer to Appendix A.3.3 for detailed validation.

A.3.1 Main theorem

Building on this Gaussian assumption, we derive bounds on the cosine similarity between potential query embeddings and both the document embedding and the mean vector.

Theorem 1. *(Concentration Inequalities for Cosine Similarities in Embedding Spaces) Let* $\mathbf{q} \sim \mathcal{N}(\mu, \Sigma)$ *denote a random vector representing the distribution of potential queries of document* d *in the unit sphere embedding space, where* $\mu \in \mathbb{R}^r$ *is the mean vector and* $\Sigma \in \mathbb{R}^{r\times r}$ *is the covariance matrix. Let* $\mathbf{d} \in \mathbb{R}^r$ *be the embedding of document* d*, and let* θ *be the angle between* μ *and* $\mathbf{d}$ *such that* $\cos(\theta) = \mu^\top \mathbf{d}$*. Assume that both* μ *and* $\mathbf{d}$ *are unit vectors. Then, the following properties hold:*

1. *The concentration inequality for the cosine similarity measure of* $\mathbf{q}$ *with* $\mathbf{d}$ *:*

$$\mathbb{P}\left(\left|\mathbf{q}^\top \mathbf{d} - \cos(\theta)\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\mathbf{d}^\top \Sigma \mathbf{d}}\right). \tag{5}$$

2. *The concentration inequality for the cosine similarity measure of* $\mathbf{q}$ *with* μ *:*

$$\mathbb{P}\left(\left|\mathbf{q}^\top \mu - 1\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\mu^\top \Sigma \mu}\right). \tag{6}$$

3. *Non-Negativity of the Difference in Similarities:*

$$\mathbf{q}\top\mu - \mathbf{q}\top\mathbf{d} = \mathbf{q}^\top\mu(1 - \cos(\theta)) > 0. \tag{7}$$

Remark 1. The theorem provides a robust theoretical foundation for the QAEncoder's capabilities in computing the similarity between documents and queries. These concentration inequalities in Equation 5 and Equation 6 show that cosine similarities between $\mathbf{q}$, $\mathbf{d}$, and μ are concentrated around their expected values. Inequality 7 indicates that the similarity between the $\mathbf{q}$ and μ is always greater than between $\mathbf{q}$ and $\mathbf{d}$. This confirms that using the mean vector as a projection in QAEncoder better captures the semantic relationship between queries and documents. This theoretical result aligns with the experimental findings in Fig. 1, further validating QAEncoder's effectiveness.

A.3.2 Proof of Similarity Bounds

Proof. Given the setup where $\mathbf{q} \sim \mathcal{N}(\mu, \Sigma)$ is an r-dimensional Gaussian random vector with mean μ and covariance Σ, and the angle between another unit vector $\mathbf{d}$ and μ is θ.

The cosine of the angle between $\mathbf{q}$ and $\mathbf{d}$ is given by:

$$\cos(\phi_{qd}) = \mathbf{q}^\top \mathbf{d}.$$

Since $\mathbf{q}$ is Gaussian, based on Lemma 1,the inner product $\mathbf{q}^\top \mathbf{d}$ is a linear transformation of $\mathbf{q}$ and hence is a normal distribution with mean $\mu^\top \mathbf{d}$ and variance $\mathbf{d}^\top \Sigma \mathbf{d}$. Given that μ and $\mathbf{d}$ are unit vectors and the angle between μ and $\mathbf{d}$ is θ, we have:

$$\mu^\top \mathbf{d} = \cos(\theta).$$

Thus, $\mathbf{q}^\top \mathbf{d}$ can be approximated as:

$$\mathbf{q}^\top \mathbf{d} \sim \mathcal{N}(\cos(\theta), \mathbf{d}^\top \Sigma \mathbf{d}).$$

The concentration inequality for the cosine value between $\mathbf{q}$ and $\mathbf{d}$ follows from Hoeffding's inequality for zero-mean sub-Gaussian random variables, which can be expressed as:

$$\mathbb{P}\left(\left|\mathbf{q}^\top \mathbf{d} - \cos(\theta)\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\mathbf{d}^\top \Sigma \mathbf{d}}\right).$$

Similarly, the cosine of the angle between between $\mathbf{q}$ and μ can be expressed as:

$$\cos(\phi_{q\mu}) = \mathbf{q}^\top \mu.$$

Given that $\mathbf{q} \sim \mathcal{N}(\mu, \Sigma)$, based on Lemma 1, the inner product $\mathbf{q}^\top \mu$, representing the cosine of the angle between $\mathbf{q}$ and μ is a linear transformation of a Gaussian random vector with mean:

$$\mathbb{E}[\mathbf{q}^\top \mu] = \mu^\top \mu = 1,$$

since μ is a unit vector, and variance:

$$\mathrm{Var}[\mathbf{q}^\top \mu] = \mu^\top \Sigma \mu.$$

Thus, $\mathbf{q}^\top \mu$ can be approximated as:

$$\mathbf{q}^\top \mu \sim \mathcal{N}(1, \mu^\top \Sigma \mu).$$

Applying the similar Hoeffding's inequality, the concentration inequality can be derived similarly:

$$\mathbb{P}\left(\left|\mathbf{q}^\top \mu - 1\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\mu^\top \Sigma \mu}\right).$$

Notably, we observe that:

$$\mathbf{q}^\top \mathbf{d} = (\mathbf{q}^\top \mu)(\mu^\top \mathbf{d}) = (\mathbf{q}^\top \mu)\cos(\theta).$$

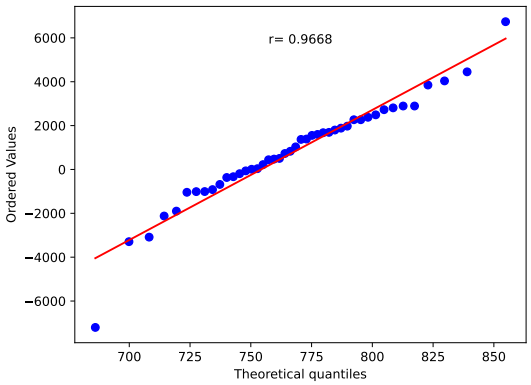


Figure 10: Q-Q plot against chi-squared distribution.

Therefore, comparing $\mathbf{q}^\top \mu$ and $\mathbf{q}^\top \mathbf{d}$, we find:

$$\mathbf{q}^\top \mu - \mathbf{q}^\top \mathbf{d} = \mathbf{q}^\top \mu (1 - \cos(\theta)).$$

Since $\mathbf{q} \neq \mathbf{d}$, it follows that $\cos(\theta) < 1$. Moreover, $\mathbf{q}^\top \mu > 0$ because $\mathbf{q}$ is a Gaussian random vector centered at μ, which implies that $\mathbf{q}$ generally aligns positively with its mean μ. Given that both $1 - \cos(\theta) > 0$ and $\mathbf{q}^\top \mu > 0$, we have:

$$\mathbf{q}^\top \mu - \mathbf{q}^\top \mathbf{d} = \mathbf{q}^\top \mu (1 - \cos(\theta)) > 0.$$

□

A.3.3 Normality Test

To assess whether potential queries follow a Gaussian distribution in the embedding space,we employ two statistical tests: the Chi-Squared Q-Q Plot and the Anderson-Darling Test.

Chi-Squared Q-Q Plot Verification We employ the Chi-Squared Q-Q Plot to assess whether the squared Mahalanobis distances 1 conform to the chi-squared distribution. By leveraging Lemma 2, We compare the observed D^2 values with the theoretical quantiles of the chi-squared distribution to assess the conformity of the data to a high-dimensional Gaussian model. We conclude that close alignment of the sample points along the 45-degree reference line indicates support for the original hypothesis.

Direct applications of high-dimensional normality tests, such as the Henze-Zirkler test, often lead to Type I errors in our case. E.g., testing on 768-dimensional normal samples revealed that Henze-Zirkler test demands high ratios of sample size to dimensionality, and it particularly susceptible to Type I errors when the sample size is not sufficiently large. Hence, we opted to perform univariate normality assessments on marginal distribution instead.

Anderson-Darling Test To further evaluate the normality of marginal distributions, we conduct the Anderson-Darling test across all dimensions of the embeddings. The specific steps are as follows:

1. Null Hypothesis:
   - H_0: Each dimension's marginal distribution follows a Gaussian distribution.
   - H_1: At least one dimension's marginal distribution does not follow a Gaussian distribution.
2. Testing Results:

   Following the Anderson-Darling test, the results across all dimensions failed to reject H_0. This suggests that each dimension's marginal distribution can statistically be considered Gaussian, thereby supporting our hypothesis that potential queries conform to a high-dimensional Gaussian distribution.

In summary, through the verification of squared Mahalanobis distances using the Chi-Squared Q-Q Plot and the evaluation of marginal distributions with the Anderson-Darling test, we validate the plausibility that potential queries conform to a high-dimensional Gaussian distribution within the embedding space.

A.3.4 Some properties of Gaussian distribution

The statement and proof of our main results contain some mathematical concepts. This section introduces these concepts, covering the fundamental lemmas and definitions essential for understanding our analysis.

Lemma 1. *(Tong, 2012) Let* $\mathbf{x}$ *follow a multivariate Gaussian distribution:*

$$\mathbf{x} \sim \mathcal{N}(\mu, \Sigma),$$

where μ *is the mean vector and* Σ *is the covariance matrix. For any linear transformation* $A\mathbf{x} + \mathbf{b}$*, the result is also multivariate normal:*

$$\mathbf{y} = A\mathbf{x} + \mathbf{b} \sim \mathcal{N}(A\mu + \mathbf{b}, A\Sigma A^\top).$$

Remark 2. The lemma shows that multivariate Gaussians remain Gaussian under linear transformations.

Definition 1. (Squared Mahalanobis Distance (McLachlan, 1999)) Assuming $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$ in a high-dimensional Gaussian distribution, the squared Mahalanobis distance D^2 is defined as:

$$D^2 = (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu).$$

Remark 3. This metric quantifies the distance between the observed value and the mean.

Lemma 2. *(Chi-Squared Distribution of Squared Mahalanobis Distance (Mardia et al., 2024)) If* $\mathbf{x}$ *is a* r*-dimensional Gaussian random variable, then the squared Mahalanobis distance* D^2 *follows a chi-squared distribution:*

$$D^2 \sim \chi^2_r,$$

where r *denotes the dimensionality of the variable.*

Remark 4. The property that the squared Mahalanobis distance follows a chi-squared distribution can be viewed as a form of dimensionality reduction. By mapping a high-dimensional Gaussian variable to a scalar that encodes its deviation from the mean, adjusted for the covariance structure, this transformation reduces the complexity of the multivariate data while preserving key statistical properties in a single distance metric.

B Query Generation Pipeline

For general purposes, we mainly utilized GPT-4o-mini and the zero-shot prompt in Appendix G for query generation on the latest datasets. For extra experiments on robustness, we also adopt Claude-3-5-Haiku-20241022, Qwen-Turbo-2024-11-01 and Deepseek-V3 as the query predictors. The temperature value is 0.95 and the frequency penalty is 0.1, making a balance between diversity and validity. That is, with a higher temperature value, the generated questions can become too divergent to be answered or supported by the documents.

This pipeline prompt is common and effective for automatic question generation (Zhang et al., 2025; Maity and Deroy, 2024). We demonstrate the diversity of generated queries and the robustness of QAEncoder representations.

B.1 The Diversity of Predicted Query

Statistics Results Zhang et al. (2025) adopted almost the same prompt and generation process. It showed the diversity and quality of generated queries from six dimensions including *question types, question length, context coverage, answerability, uncommonness, and answer length*. We excerpt the table of question types, showing the diversity of generated queries.

	TriviaQA	HotpotQA	Llama	GPT
T1 Identity/Attribution	34.2	39.7	7.5	15.7
T2 General Knowledge	34.5	15.0	7.3	12.3
T3 Location	12.2	14.3	3.1	4.3
T4 Classification/Categorization	4.3	2.7	2.1	1.5
T5 Specific Fact/Figure	10.5	9.5	18.7	24.1
T6 Comparison/Selection	0.1	6.7	1.0	0.6
T7 Verification/Affirmation	0.1	6.5	0.1	0.2
T8 Descriptive/Characterization	3.0	1.5	43.8	28.7
T9 Event/Outcome	0.2	0.8	14.7	10.1
T10 Sequential/Ordering/Causation	0.9	3.2	1.7	2.4
Others	0.0	0.0	0.0	0.1

Table 6: Percentage of different question types across different datasets

Case Study To provide a general feeling of diversity, we show the predicted queries by GPT and Claude. For two constructed news benchmarks, the test query is sampled without replacement from GPT's results; for BEIR, the test queries are from original benchmarks.

Case Study of Diversity

```
- Document: On July 20, 1969,
    Neil Armstrong landed on the
    moon to accomplish the Apollo
    11 mission, achieved by NASA's
    planning and the Saturn V
    rocket.
- GPT: ['1. What significant
    event took place on July 20,
    1969?', '2. Who was the first
    person to set foot on the
    moon?', '3. Which organization
    was responsible for planning
    the Apollo 11 mission?', '4.
    What was the name of the
    rocket used to accomplish the
    Apollo 11 mission?', '5. What
    achievement was accomplished
    by Neil Armstrong during the
    Apollo 11 mission?']
- Claude: ['1. On what date did
    Neil Armstrong land on the
    moon?', '2. What was the name
    of the mission during which
    Neil Armstrong landed on the
    moon?', '3. Which organization
    planned the moon landing?',
    '4. What type of rocket was
    used to achieve the moon
    landing?', '5. Who was the
    astronaut who landed on the
    moon during the Apollo 11
    mission?']
```

C Cost-Effectiveness

Briefly, QAEncoder is cost-effective and efficient.

- Firstly, the cost mainly comes from the query generation process during indexing phase. As shown in Fig. 11, 10 queries with an average length of 6-15 tokens are typically sufficient for ordinary documents.
- Secondly, GPT-4o-mini is currently the cheapest generative model by OpenAI, priced at $0.075 per million input tokens and $0.300 per million output tokens. The most recent Qwen2.5-Turbo by Alibaba is even 3.6 times cheaper than GPT-4o-mini, which outperforms GPT-4 on long-context and mirrors GPT-4o-mini on short-sequence (Qwen Development Team, 2024b). Hence, about 0.1 million documents can be processed within 1 dollar via API call ($\frac{1\text{M}*3.6}{0.3*(10*10)} \approx 0.1\text{M}$). The prices will continually decrease as AI develops.
- Thirdly, the initial query-centric work for sparse retrievers, DocT5Query, demonstrates T5-base with 0.2B parameters is sufficient for query generation, while no improvement are gained with larger models (Cheriton, 2019). We also confirmed Qwen2.5-0.5B-Instruct as a good query generator in our business implementation. We choose GPT-4o-mini just given its out-of-the-box and comprehensively multilingual support for academic research.
- Finally, in our data flow settings (batch size=1), 2-3 documents can be processed per second with Qwen2.5-0.5B-Instruct + vLLM + BF16 on a single NVIDIA A100 80GB (Qwen Development Team, 2024a). For batch processing with higher GPU utility, DocT5Query reports sampling 5 queries per document for 8.8M MSMARCO documents requires approximately 40 hours on a single Google TPU v3, costing only $96 USD (40 hours × $2.40 USD/hour) in 2019 (Cheriton, 2019).

D Dataset Details

- The BEIR benchmark is a meticulously curated collection of 19 datasets, designed to comprehensively evaluate the generalization capabilities of information retrieval (IR) models across a wide range of heterogeneous tasks. Among the 19 datasets, 15 are publicly available and are used for evaluation in our experiments. The 15 datasets selected from BEIR encompass diverse domains: MSMARCO (Nguyen et al., 2016), TREC-COVID(TRECC.) (Voorhees et al., 2021), NFCorpus (Boteva et al., 2016), Natural Questions (NQ) (Kwiatkowski et al., 2019), HotpotQA (Yang et al., 2018), FiQA18 (Maia et al., 2018), ArguAna (Wachsmuth et al., 2018), Touche20 (Bondarenko et al., 2022), CQADupStack(CQADup.) (Hoogeveen et al., 2015), Quora, DBPedia (Hasibi et al., 2017), SciDocs (Cohan et al., 2020), Fever (Thorne et al., 2018), Climate-Fever(ClimateFe.) (Diggelmann et al., 2020), and SciFact (Wadden et al., 2020).
- CRUD-RAG is a benchmark specifically designed for evaluating RAG systems, incorporating the latest high-quality news data that were not included in the training phase of the language models. It comprises more than 80K news articles sourced from prominent Chinese news websites, all published after July 2023. From the set of queries generated by GPT-4o-mini for each document, we randomly sample one to serve as the test query. The original news is designated as the evidence documents for recall evaluation, ensuring queries are associated with exactly relevant documents.[2]
- FIGNEWS is a multilingual news post dataset designed to examine bias and propaganda within news articles across different languages. It consists of 15,000 publicly available news posts collected from verified blue-check accounts between October 7, 2023, and January 31, 2024. The dataset includes posts in five languages—English, Arabic, Hebrew, French, and Hindi—distributed evenly across 15 batches, each containing 1,000 posts. Each batch consists of 200 posts for each language. Similar to CRUD-RAG, we randomly sample one predicted query generated by GPT-4o-mini for each document and use the original news as the evidence documents for recall evaluation.

[2]LlamaIndex adopts the same common practice and provides templated workflow, i.e. generating queries for recall and rerank evaluation. See the website.

E Metric Details

- Mean Reciprocal Rank (MRR): Mean Reciprocal Rank: is a statistic measure used to evaluate the effectiveness of a retrieval system by calculating the reciprocal of the rank at which the first relevant result appears. The mathematical formulation is:

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\text{rank}_i}$$

Where: - $|Q|$ is the number of queries. - rank_i is the rank position of the first relevant document for the i-th query.

- Normalized Discounted Cumulative Gain (NDCG): Normalized Discounted Cumulative Gain is a measure of ranking quality that takes into account the positions of the relevant documents. It is based on the concept of discounting the relevance of documents based on their position in the result list. The mathematical formulation is:

$$\text{NDCG} = \frac{DCG_p}{IDCG_p},$$

Where: - DCG_p is the Discounted Cumulative Gain at position p. - $IDCG_p$ is the Ideal Discounted Cumulative Gain at position p, which is the DCG score of the perfect ranking.

The Discounted Cumulative Gain at position p is given by:

$$DCG_p = \sum_{i=1}^{p}\frac{2^{\text{rel}_i}-1}{\log_2(i+1)},$$

Where: - rel_i is the relevance score of the document at rank i.

The Ideal Discounted Cumulative Gain $IDCG_p$ is computed in the same way as DCG_p, except that the documents are ideally sorted by relevance.

F Baseline Details

- BM25 (Robertson and Zaragoza, 2009) is the traditional lexical retriever based on term relevance and frequency, regarded as the most popular variation of TF-IDF.
- DocT5Query (Cheriton, 2019) appends generated queries to the document before building the inverted index of BM25.
- BGE models (Xiao et al., 2024)by BAAI, Jina models (Günther et al., 2023) by Jina AI, E5 models (Wang et al., 2023a) by Microsoft, GTE models (Zhang et al., 2024b) by Alibaba-NLP are the most advanced embedding models, featuring multilingual understanding and task-specific instruction tuning capabilities. We choose both vanilla encoders and QA-specific instruction-tuned encoders for test.
- Contriever models (Izacard et al., 2021) are developed by Facebook Research, including contriever, mcontriever, contriever-msmacro and mcontriever-msmacro. mcontriever serves as the multilingual version of contriever. contriever-msmacro and mcontriever-msmacro are further fine-tuned on the MS-MACRO dataset for bridging the document-query gap.
- BCEmbedding models (NetEase Youdao, 2023), developed by NetEase Youdao, are bilingual and crosslingual embedding models in English and Chinese. BCEmbedding serves as the cornerstone of Youdao's RAG-based QA system, QAnything, an open-source project widely integrated in commercial products like Youdao Speed Reading and Youdao Translation. We choose bce-embedding-base for test.
- Text2Vec models (Xu, 2023) is a popular open-source project that implements Word2Vec (Mikolov et al., 2013), RankBM25, BERT (Devlin et al., 2019), Sentence-BERT (Reimers and Gurevych, 2019), CoSENT and other text representation models. We test its most prominent model, text2vec-base-multilingual, which supports multiple languages, including German, English, Spanish, French, Italian, Dutch, Polish, Portuguese, Russian, and Chinese.
- DPR models (Karpukhin et al., 2020) by Facebook adopt a bi-encoder architecture. DPR models fine-tuned BERT on pairs of questions and passages without additional pretraining, achieving superior performance compared to traditional methods like BM25.

- RePAQ models (Lewis et al., 2021) are dense retrievers trained on PAQ (Lewis et al., 2021), a large-scale synthetic corpus of question–answer pairs generated from web texts. We evaluate retriever-multi-base-256, a multilingual variant designed for efficient passage retrieval across multiple languages.

- Quora DistilBERT models (Reimers and Gurevych, 2019), released by the Sentence-Transformers project, are sentence embedding models based on DistilBERT and fine-tuned on the Quora Duplicate Questions dataset (Thakur et al., 2021) for semantic similarity and duplicate question detection. We evaluate both quora-distilbert-base and quora-distilbert-multilingual, the latter being a multilingual variant trained on parallel corpora covering over 50 languages.

- Training-based approaches mainly include two types: fine-tuning on QA datasets (domain adaptation) and fine-tuning on multi-task instruction datasets. For fine-tuning on QA datasets, we choose mcontriever-msmarco for test, an enhanced variant of the mcontriever model that has been fine-tuned on the MSMARCO. The second category involves fine-tuning models on multi-task instruction datasets, where distinct prompt prefixes are appended to the input text, enabling the model to effectively differentiate between various tasks. In this category, we test the multilingual-e5-large-instruct (Wang et al., 2024) developed by Microsoft, which leverages synthetic instruction data (Wang et al., 2023a) for fine-tuning. To unveil the catastrophic forgetting issue of training-based methods, we incorporate GPL (Wang et al., 2021), which predicted queries, mines hard negative samples, and distills the re-ranker for unsupervised domain adaptation. The method utilize pseudo-queries generated from the target domain as supervision for contrastive learning. Following the original settings, we train the GPL models for 140,000 steps with a batch size of 32.

- Document-centric methods instruct LLMs to generate a pseudo-document for each query. The pseudo-document aims to capture relevant information but does not correspond to a real document and may contain inaccuracies and hallucinations. Subsequently,the pseudo-document is encoded, and its embedding is utilized to retrieve similar real documents based on vector similarity. We choose HyDE (Gao et al., 2023), Query2Doc (Wang et al., 2023b), and QA-RAG (Kim and Min, 2024) for test. Hyde generates multiple pseudo-documents and fuses their embeddings by mean pooling for retrieval. Similarly, Query2Doc concatenates the query with the generated pseudo-document and performs dense retrieval using the embedding of the combined text. QA-RAG (Kim and Min, 2024) enhances retrieval through a two-way mechanism that utilizes both user query and pseudo-documents for respective retrieval.

G Instruction Templates

We employ specialized prompts to instruct GPT-4o-mini as the question and pseudo-document generator respectively. Only the English prompts are presented due to LaTeX compilation issues with non-English languages. For BEIR benchmarks, declarative words can serve as valid user queries. We adopt the few-shot in-context learning, which enables the model to generate declarative queries that not only resemble natural inputs but also serve a role analogous to document fingerprints.

Question Generator Prompt

```
Context information is below.
---------------------
[Document]
---------------------
Given the context information and
    not prior knowledge, generate
    only questions based on the
    below query.
You are a Teacher/Professor. Your
    task is to setup [Number of
    Questions] questions for an
    upcoming quiz/examination. The
    questions should be diverse in
    nature across the document.
    Restrict the questions to the
    information provided, and
    avoid ambiguous references.

Output Format:
```json
[
    "1. question",
    "2. question",
    ...
]
```
```

Pseudo-Document Generator Prompt

```
Please write a passage to answer
    the question.
Question: [Question]
Output Format:
```json
{
    "passage": ""
}
```
```

H Hyperparameter Selection

As shown in the main body, QAE_{emb} maintains competitive performance with single hyperparameter. Hence, QAE_{emb} is recommended for accelerating HP search.

- Firstly, we believe that the optimal hyperparameters are primarily influenced by the inherent characteristics of the embedding model, i.e. the geometric property of embedding space. Therefore, a one-turn search should be sufficient for a given embedding model. That's why we optimize hyperparameters simultaneously across multiple datasets.
- Secondly, the one-turn search can also be accelerated under our framework. Indeed, as Fig. 4 shows, the performance of QAE_{emb} empirically follows a consistent trend across various models and datasets: it initially rises and then falls as α increases, peaking between 0.3 and 0.6. This unimodal phenomenon enables ternary search with logarithmic trails rather than brute-force search.
- Finally, the property of datasets also slightly influences the optimal hyperparameters. Specifically, the optimal α for classical datasets is marginally lower than that for latest datasets (refer to Tables 7 and 8 for details). Therefore, selecting the optimal α based on classical datasets represents a cautious and robust strategy, ensuring consistent improvement across both classical and latest datasets.

I Challenges of Existing Methods

I.1 Challenges of $\text{QAE}_{\text{naive}}$

C1. Expanded Index Size. Storing all QA pairs significantly increases the index size, leading to a substantial expansion in storage requirements, especially problematic for large-scale corpora.

C2. Prolonged Retrieval Times. The index expansion also results in extended retrieval times. For dense retrievers, the expanded index size can result in linearly increased search time in both exhaustive and non-exhaustive search (Douze et al., 2024) and hurt the recall performance in non-exhaustive case (Zhao et al., 2023).

C3. Limited Query Handling. Although storing QA pairs individually can address predicted queries, this approach lacks robustness when confronted with the wide-ranging and diverse nature of potential queries (Alting von Geusau and Bloem, 2020).

I.2 Challenges of Document-centric Methods

Document-centric methods such as HyDE and Query2Doc suffer from not only computation overhead at inference time but also hallucination of pseudo-document generation. Specifically, for user queries, these methods generate pseudo-documents as retrieval queries. However, pseudo-documents suffer from hallucination, especially with rapidly updated knowledge. A concrete example hallucination can be:

Pseudo-Document Generator Prompt

```
- User Query: Who won the 2024
    Abel Prize?
- Pseudo Document (Hallucinatory
    Retrieval Query): The 2024
    Abel Prize was awarded to Abel
    for proving that it is
    impossible to solve the
    general equation of the fifth
    degree using radicals.
- Retrieved Document: Abel's most
    famous single result is the
    first complete proof
    demonstrating the
    impossibility of solving the
    general quintic equation in
    radicals. This question was
    one of the outstanding open
    problems of his day, and had
    been unresolved for over 250
    years.[2] He was also an
    innovator in the field of
    elliptic functions and the
    discoverer of Abelian
    functions.
- LLM Answer: Abel
- Groundtruth: Michel Talagrand
```

J More Figures and Tables

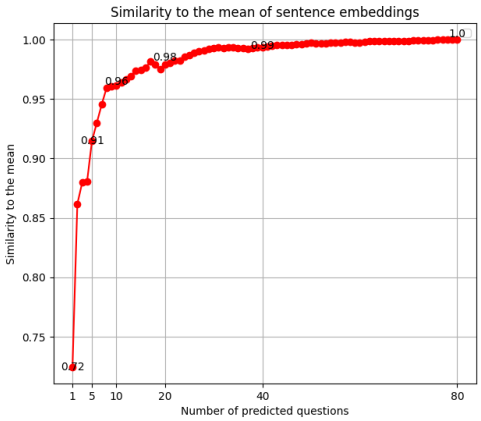


Figure 11: The convergence speed for Monte Carlo estimation of QAE_{base}, n denotes the number of prediction queries. For documents with a length greater than 150 words from MSMARCO datasets, generating 10 queries exhibits a similarity score of 0.96 compared to generating 80 queries. Besides, the document fingerprint strategies introduces the hyperparameter α to QAE_{base}, further reducing the variance.

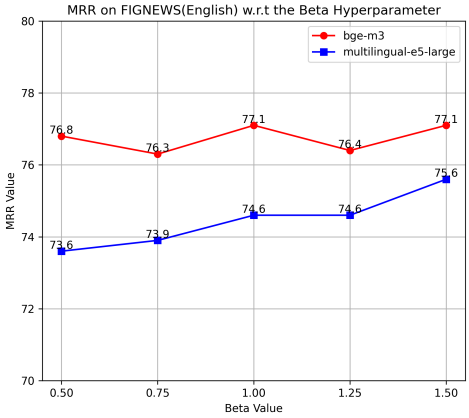


Figure 12: Ablation on β hyperparameter for QAE_{txt} on FIGNEWS(English) dataset.

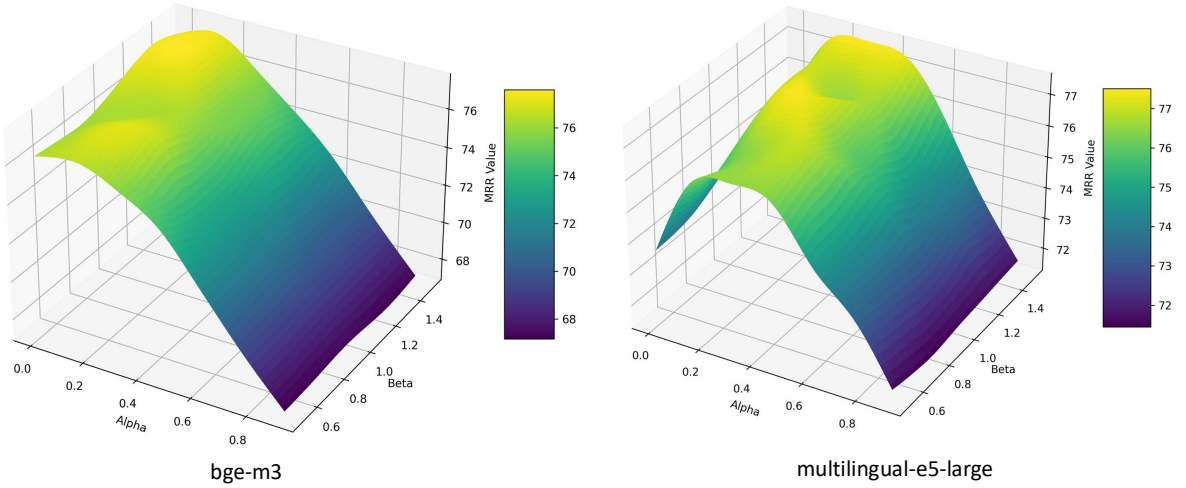


Figure 13: Ablation on α and β hyperparameters for QAE_{hyb} on FIGNEWS(English) dataset.

Model	Method	ArguAna	ClimateF.	CQADups.	DBPedia	FEVER	FiQA18	HotpotQA	MSMARCO	NFCorpus	NQ	Quora	SciDocs	SciFact	Touche20	TRECC.
Sparse																
BM25	-	31.5	**21.3**	29.9	31.3	**75.3**	23.6	**60.3**	22.8	**32.5**	32.9	78.9	15.8	66.5	**36.7**	65.6
DocT5Query	-	**34.9**	20.1	**32.5**	**33.1**	71.4	**29.1**	58.0	**33.8**	**32.8**	**39.9**	**80.2**	**16.2**	**67.5**	34.7	**71.3**
Dense																
dpr	-	17.5	14.8	15.3	26.3	56.2	11.2	39.1	17.7	18.9	47.4	24.8	7.7	31.8	13.1	33.2
	QAE_{emb}, $\alpha = 0.45$	**29.6**	**27.6**	**22.3**	**32.4**	**70.9**	**20.3**	**43.4**	**25.4**	**23.9**	**52.7**	**28.9**	**13.5**	**40.4**	**22.9**	**48.0**
contriever	-	37.9	15.5	28.4	29.2	68.2	24.5	48.1	20.6	31.7	25.4	83.5	14.9	64.9	19.3	27.4
	QAE_{emb}, $\alpha = 0.45$	**47.1**	**20.2**	**33.5**	**33.7**	**73.1**	**30.4**	**50.3**	**26.1**	**37.9**	**28.1**	**86.2**	**17.3**	**70.2**	**25.2**	**45.2**
contriever-msmarco	-	44.6	23.7	34.5	41.3	75.8	32.9	63.8	**40.7**	32.8	49.8	86.5	16.5	67.7	20.4	59.6
	QAE_{hyb}, $\alpha = 0.3,\ \beta = 0.75$	**53.9**	**25.5**	**38.2**	**44.9**	**82.0**	**37.9**	**65.3**	39.9	**37.6**	**52.7**	**89.9**	**18.7**	**73.6**	**26.3**	**70.6**
bge-large-en-v1.5	-	63.5	36.6	42.2	44.1	87.2	45.0	**74.1**	**42.5**	38.1	55.0	89.1	22.6	74.6	24.8	74.8
	QAE_{hyb}, $\alpha = 0.15,\ \beta = 0.5$	**68.8**	**38.2**	**45.6**	**44.8**	**91.5**	**48.2**	73.9	41.2	**41.7**	**56.4**	**90.1**	**25.6**	**78.9**	**28.1**	**78.2**
multilingual-e5-large	-	54.4	25.7	39.7	41.3	**82.8**	43.8	**71.2**	**43.7**	34.0	64.1	**88.2**	17.5	70.4	23.1	71.2
	QAE_{hyb}, $\alpha = 0.15,\ \beta = 1.5$	**61.1**	**28.4**	**44.3**	**42.9**	82.1	**45.1**	69.9	43.0	**35.5**	**65.3**	88.0	**20.8**	**73.9**	**26.3**	**75.1**
e5-large-v2	-	46.4	22.2	37.9	44.0	82.8	41.1	73.1	**43.5**	37.1	63.4	86.8	20.5	72.2	20.7	66.5
	QAE_{hyb}, $\alpha = 0.3,\ \beta = 1.0$	**55.1**	**25.3**	**41.2**	**45.5**	**86.5**	**43.4**	**73.9**	42.8	**39.8**	**64.8**	**89.5**	**23.3**	**75.3**	**23.8**	**74.2**
gte-base-en-v1.5	-	63.5	40.4	39.5	**39.9**	**94.8**	48.7	**67.8**	**42.6**	35.9	**53.0**	**88.4**	21.9	76.8	25.2	73.1
	QAE_{hyb}, $\alpha = 0.3,\ \beta = 0.5$	**68.2**	**43.2**	**43.7**	39.6	94.2	**51.6**	66.7	41.9	**38.7**	52.3	88.2	**24.8**	**80.3**	**29.3**	**77.5**
jina-embeddings-v2-small-en	-	46.7	24.0	38.0	32.7	68.0	33.4	**56.5**	37.3	30.4	**51.6**	87.2	18.6	63.9	23.5	65.2
	QAE_{hyb}, $\alpha = 0.15,\ \beta = 0.5$	**55.3**	**27.8**	**42.3**	**35.6**	**74.3**	**36.5**	55.7	**39.5**	**32.7**	51.1	**89.9**	**21.2**	**67.2**	**27.2**	**73.6**

Table 7: Complete retrieval performance across fifteen classical datasets in BEIR (NDCG@10). Higher is better, with the best one is bolded. Hyperparameters including QAEncoder variants and weight terms $\alpha,\ \beta$ are optimized simultaneously. '-' denotes default or null values.

Model	Method	FIGNEWS(English)		FIGNEWS(Arabic)		CRUD-RAG(Chinese)		FIGNEWS(French)		FIGNEWS(Hindi)		FIGNEWS(Hebrew)	
		MRR@10	NDCG@10	MRR@10	NDCG@10	MRR@10	NDCG@10	MRR@10	NDCG@10	MRR@10	NDCG@10	MRR@10	NDCG@10
bge-m3	-	74.4	78.7	77.8	80.9	47.5	48.6	73.5	77.4	58.6	64.4	78.7	81.4
	QAE_{txt}, $\beta = 1.5$	**77.2**	**81**	**80.2**	**83.1**	**51.4**	**52.5**	**76.9**	**80.3**	**62.7**	**67.8**	**80**	**82.8**
multilingual-e5-small	-	71	75.1	74.1	77.4	44.6	46.0	66.8	70.4	52.5	57.8	72.9	76.5
	QAE_{hyb}, $\alpha = 0.3,\ \beta = 0.5$	**74.6**	**78.5**	**78.9**	**81.6**	**50.6**	**51.6**	**74.2**	**77.9**	**59.6**	**64.6**	**77.5**	**80.4**
multilingual-e5-base	-	74.8	78.1	72.3	76	47.0	48.2	71.2	75.1	57.8	62.9	72.6	75.8
	QAE_{emb}, $\alpha = 0.3$	**77.6**	**81.3**	**77.2**	**80.3**	**51.2**	**52.3**	**76.7**	**80**	**61.5**	**66.5**	**77.7**	**80.5**
multilingual-e5-large	-	73.9	77.8	76.7	80.2	46.9	48.3	70.6	74.5	53	59.2	73.9	77.4
	QAE_{hyb}, $\alpha = 0.15,\ \beta = 1.25$	**77.1**	**80.6**	**82.2**	**85.1**	**51.5**	**52.7**	**77.4**	**80.9**	**60.6**	**65.9**	**77.7**	**81**
gte-multilingual-base	-	65.5	70.4	73.4	76.8	45.3	46.8	63	67.4	52.4	58.6	66	69.6
	QAE_{hyb}, $\alpha = 0.15,\ \beta = 1.5$	**75.5**	**79.5**	**76.2**	**79.1**	**49.4**	**51.0**	**66.9**	**71.6**	**56**	**61.9**	**74**	**77.6**
mcontriever	-	32.9	36.7	40.3	44.7	39.2	41.6	35	39.3	27	31.8	49.5	54.4
	QAE_{hyb}, $\alpha = 0.45,\ \beta = 1.25$	**61.4**	**65.9**	**68.3**	**72.1**	**51.3**	**52.4**	**64.7**	**69.1**	**50.6**	**56.4**	**70.1**	**73.8**
bce-embedding-base-v1	-	59.1	63.8	-	-	42.0	44.0	-	-	-	-	-	-
	QAE_{hyb}, $\alpha = 0.3,\ \beta = 0.5$	**66.8**	**71.1**	-	-	**49.7**	**51.0**	-	-	-	-	-	-
text2vec-base-multilingual	-	38.7	43.6	27.8	31.9	9.7	10.6	33.7	38.6	15.6	19.9	12.6	15.5
	QAE_{emb}, $\alpha = 0.75$	**55.4**	**59.9**	**51.5**	**55.4**	**32.1**	**34.1**	**49.3**	**54.4**	**36.2**	**41.5**	**47.1**	**52.1**
Ablation													
bge-m3	-	74.4	78.7	77.8	80.9	47.5	48.6	73.5	77.4	58.6	64.4	78.7	81.4
	QAE_{emb}, $\alpha = 0.3$	76.4	80.5	80.1	82.9	51.3	52.4	75	78.5	61	66	79	82
	QAE_{txt}, $\beta = 1.5$	77.2	81	80.2	83.1	51.4	52.5	76.9	80.3	62.7	67.8	80	82.8
	QAE_{hyb}, $\alpha = 0.15,\ \beta = 1.5$	77.4	81.1	80.6	83.4	51.7	52.7	76.5	80	61.7	66.8	79.8	82.7
multilingual-e5-small	-	71	75.1	74.1	77.4	44.6	46.0	66.8	70.4	52.5	57.8	72.9	76.5
	QAE_{emb}, $\alpha = 0.45$	74.7	78.5	77	79.8	50.8	51.9	73.6	77.1	58	63.2	77.1	80.1
	QAE_{txt}, $\beta = 1.0$	73.2	77.2	79.2	81.9	49.1	50.5	70.6	74.7	58.7	63.7	77.1	80.2
	QAE_{hyb}, $\alpha = 0.3,\ \beta = 0.5$	74.6	78.5	78.9	81.6	50.6	51.6	74.2	77.9	59.6	64.6	77.5	80.4
multilingual-e5-base	-	74.8	78.1	72.3	76	47.0	48.2	71.2	75.1	57.8	62.9	72.6	75.8
	QAE_{emb}, $\alpha = 0.3$	77.6	81.3	77.2	80.3	51.2	52.3	76.7	80	61.5	66.5	77.7	80.5
	QAE_{txt}, $\beta = 0.75$	74.7	78.8	76.4	79.7	50.6	51.8	72.3	76.4	62.5	67.8	77	79.9
	QAE_{hyb}, $\alpha = 0.3,\ \beta = 0.5$	76.6	80.4	77.4	80.6	51.5	52.5	75.4	79.1	62.7	67.7	77.2	80.2
multilingual-e5-large	-	73.9	77.8	76.7	80.2	46.9	48.3	70.6	74.5	53	59.2	73.9	77.4
	QAE_{emb}, $\alpha = 0.45$	77.9	81.4	79.8	83	51.9	52.9	77	80	58.3	63.5	78.1	81.2
	QAE_{txt}, $\beta = 1.5$	75.6	79.2	80.9	84.1	51.0	52.3	76	79.4	60	65.5	77	80.2
	QAE_{hyb}, $\alpha = 0.15,\ \beta = 1.25$	77.1	80.6	82.2	85.1	51.5	52.7	77.4	80.9	60.6	65.9	77.7	81
gte-multilingual-base	-	65.5	70.4	73.4	76.8	45.3	46.8	63	67.4	52.4	58.6	66	69.6
	QAE_{emb}, $\alpha = 0.45$	69.1	73.6	76.9	79.6	50.5	51.7	67.1	71.2	53	58.5	72.4	75.8
	QAE_{txt}, $\beta = 1.5$	75.7	79.7	75.9	78.9	48.8	50.5	66.1	70.7	56.8	62.5	72.8	76.5
	QAE_{hyb}, $\alpha = 0.15,\ \beta = 1.5$	75.5	79.5	76.2	79.1	49.4	51.0	66.9	71.6	56	61.9	74	77.6
mcontriever	-	32.9	36.7	40.3	44.7	39.2	41.6	35	39.3	27	31.8	49.5	54.4
	QAE_{emb}, $\alpha = 0.6$	58.8	64.2	67.2	71.3	51.0	52.1	62.7	66.9	50.4	55.7	69.7	73.2
	QAE_{txt}, $\beta = 1.5$	49.2	54.2	59.5	63.9	46.2	48.2	54.1	58.9	45.1	50.2	60.8	64.6
	QAE_{hyb}, $\alpha = 0.45,\ \beta = 1.25$	61.4	65.9	68.3	72.1	51.3	52.4	64.7	69.1	50.6	56.4	70.1	73.8
bce-embedding-base-v1	-	59.1	63.8	-	-	42.0	44.0	-	-	-	-	-	-
	QAE_{emb}, $\alpha = 0.45$	66.8	71.3	-	-	49.3	50.6	-	-	-	-	-	-
	QAE_{txt}, $\beta = 1.5$	64.3	68.7	-	-	47.6	49.2	-	-	-	-	-	-
	QAE_{hyb}, $\alpha = 0.3,\ \beta = 0.5$	66.8	71.1	-	-	49.7	51.0	-	-	-	-	-	-
text2vec-base-multilingual	-	38.7	43.6	27.8	31.9	9.7	10.6	33.7	38.6	15.6	19.9	12.6	15.5
	QAE_{emb}, $\alpha = 0.75$	55.4	59.9	51.5	55.4	32.1	34.1	49.3	54.4	36.2	41.5	47.1	52.1
	QAE_{txt}, $\beta = 1.5$	46.2	50.8	35.5	39.4	12.3	13.6	43.3	47.7	21.3	25.2	18.8	22.1
	QAE_{hyb}, $\alpha = 0.75,\ \beta = 0.5$	55.1	59.7	50.4	54.4	33.4	35.1	49.1	53.8	34.5	39.6	47.1	51.5

Table 8: Comprehensive retrieval performance on the latest datasets FIGNEWS and CRUD-RAG (**Top-k = 10**). Higher is better, with the best one bolded. Hyperparameters including QAEncoder variants and weight terms $\alpha,\ \beta$ are optimized simultaneously for six latest datasets. '-' denotes default or null values.

Model	Method	FIGNEWS(English)		FIGNEWS(Arabic)		CRUD-RAG(Chinese)		FIGNEWS(French)		FIGNEWS(Hindi)		FIGNEWS(Hebrew)	
		MRR@10	NDCG@10	MRR@10	NDCG@10	MRR@10	NDCG@10	MRR@10	NDCG@10	MRR@10	NDCG@10	MRR@10	NDCG@10
bg3-m3	QAE_{emb}, $\alpha = 0.3$	76.4	80.5	80.1	82.9	88.8	90.7	75	78.5	61	66	79	82
	QAE_{txt}, $\beta = 1.5$	77.2	81	80.2	83.1	89	90.9	**76.9**	**80.3**	**62.7**	**67.8**	**80**	**82.8**
	QAE_{hyb}, $\alpha = 0.15,\ \beta = 1.5$	**77.4**	**81.1**	**80.6**	**83.4**	**89.4**	**91.2**	76.5	80	61.7	66.8	79.8	82.7
	QA_{naive}, n=10	76.9	79.9	77.1	79.7	86	88	71.9	74.4	62.3	66.1	68.1	71.7
multilingual-e5-large	QAE_{emb}, $\alpha = 0.45$	**77.9**	**81.4**	79.8	83	**89.8**	**91.5**	77	80	58.3	63.5	**78.1**	**81.2**
	QAE_{txt}, $\beta = 1.5$	75.6	79.2	80.9	84.1	88.3	90.5	76	79.4	60	65.5	77	80.2
	QAE_{hyb}, $\alpha = 0.15,\ \beta = 1.25$	77.1	80.6	**82.2**	**85.1**	89.1	91.2	**77.4**	**80.9**	60.6	**65.9**	77.7	81
	QA_{naive}, n=10	77.5	80.3	76.5	79.4	85.1	87.3	70.5	73.5	**61.5**	65.6	69.4	72.2

Table 9: Complete performance comparison of QAEncoder variants on latest datasets FIGNEWS and CRUD-RAG (Top-k = 10). Higher is better, with the best one bolded. Hyperparameters are optimized simultaneously across the six latest datasets. n indicates the number of predicted queries in QA_{naive}.

Model	Method	FIGNEWS(English)		FIGNEWS(Arabic)		CRUD-RAG(Chinese)		FIGNEWS(French)		FIGNEWS(Hindi)		FIGNEWS(Hebrew)	
		MRR@10	NDCG@10	MRR@10	NDCG@10	MRR@10	NDCG@10	MRR@10	NDCG@10	MRR@10	NDCG@10	MRR@10	NDCG@10
mcontriever	-	32.9	36.7	40.3	44.7	39.2	41.6	35	39.3	27	31.8	49.5	54.4
	QAE_{hyb}, $\alpha = 0.45,\ \beta = 1.25$	61.4	65.9	68.3	72.1	51.3	52.4	64.7	69.1	50.6	56.4	70.1	73.8
	$MS^{\dagger}$	66.1	70.6	70.2	73.7	46.5	47.8	66.4	70.3	48.7	53.9	69.4	72.9
	$MS^{\dagger}$ + QAE_{hyb}, $\alpha = 0.3,\ \beta = 0.75$	**72.3**	**76.8**	**77.3**	**80.5**	51.2	52.4	**74.4**	**78**	**59**	**64.1**	**76.5**	**79.8**
	$GPL^{\dagger}$	68.3	73.6	72.1	75.18	**52.8**	**55.9**	70.8	74.3	52.6	55.9	73.9	77.1
	QA-RAG‡	31.1	34.4	42.3	46.8	43.8	45.8	31.8	35.5	29.1	33.9	49.7	55.0
	Query2Doc‡	25.1	29.0	34.5	38.9	35.7	37.2	26.0	29.1	14.2	17.2	39.9	45.0
	HyDE‡	25	27.9	35.7	41.9	36.7	38.7	25.8	28.9	11.8	14.9	42.1	47.8
multilingual-e5-large	-	73.9	77.8	76.7	80.2	46.9	48.3	70.6	74.5	53	59.2	73.9	77.4
	QAE_{hyb}, $\alpha = 0.15,\ \beta = 1.25$	**77.1**	**80.6**	**82.2**	**85.1**	51.5	52.7	**77.4**	**80.9**	**60.6**	**65.9**	77.7	81
	$INS^{\dagger}$	67	71.4	75	78.2	43.7	45.2	66	70.9	48.2	53.8	69.6	73.3
	$INS^{\dagger}$ + QAE_{hyb}, $\alpha = 0.15,\ \beta = 1.5$	75.6	79.8	80.8	83.7	51.4	52.4	75.8	79.5	59.3	65	**79.4**	**82.3**
	$GPL^{\dagger}$	75.2	78.9	79.4	82.3	**53.6**	**56.3**	73.7	76.8	57.5	62.8	74.8	78.3
	QA-RAG‡	73.3	76.5	72.8	75.6	45.8	46.5	67.2	70.2	55.5	61.3	77.7	80.4
	Query2Doc‡	63.4	68.2	66.3	72.8	42.0	43.1	60.4	65.1	48.4	53.3	62.1	66.0
	HyDE‡	63.6	68.3	68.3	74.1	42.3	43.6	58.4	63.2	45.6	51.3	65.3	69.6

Table 10: The table illustrates a comprehensive performance comparison of QAEncoder against training-based and document-centric methods on the latest datasets: FIGNEWS and CRUD-RAG (Top-k = 10). Higher is better, with the best one bolded. Hyperparameters $\alpha,\ \beta$ are optimized simultaneously across the six latest datasets. † indicates the training-based methods. $MS^{\dagger}$ represents fine-tuning on MSMARCO, i.e. the mcontriever-msmarco model; $INS^{\dagger}$ represents instruction-tuning, i.e. the multilingual-e5-large-instruct model(Wang et al., 2023a). ‡ indicates the document-centric methods.

Model	Param	FIGNEWS(English)		FIGNEWS(Arabic)		CRUD-RAG(Chinese)	
		MRR@10	NDCG@10	MRR@10	NDCG@10	MRR@10	NDCG@10
jina-embeddings-v2-small-en	-	65.4	69.9	-	-	-	-
	QAE_{hyb}, $\alpha = 0.15,\ \beta = 1.25$	**72.6**	**76.7**	-	-	**-**	**-**
jina-embeddings-v2-base-zh	-	-	-	-	-	41.0	43.1
	QAE_{hyb}, $\alpha = 0.75,\ \beta = 0.75$	**-**	**-**	**-**	**-**	**47.5**	**48.5**
jina-embeddings-v2-base-en	-	65.3	69.4	-	-	-	-
	QAE_{hyb}, $\alpha = 0.3,\ \beta = 0.5$	**72.4**	**76.5**	**-**	**-**	**-**	**-**
gte-base-en-v1.5	-	65.6	70.2	-	-	-	-
	QAE_{hyb}, $\alpha = 0.15,\ \beta = 1.5$	**71.3**	**75.6**	**-**	**-**	**-**	**-**
contriever	-	49.6	54.5	-	-	-	-
	QAE_{hyb}, $\alpha = 0.45,\ \beta = 1.25$	**70.7**	**74.7**	**-**	**-**	**-**	**-**
bge-large-zh-v1.5	-	-	-	-	-	42.0	43.7
	QAE_{hyb}, $\alpha = 0.45,\ \beta = 1.25$	**-**	**-**	-	-	**48.6**	**49.6**
bge-large-zh	-	-	-	-	-	40.6	42.5
	QAE_{hyb}, $\alpha = 0.45,\ \beta = 1.5$	**-**	**-**	**-**	**-**	**48.4**	**49.4**
bge-large-en-v1.5	-	66.4	71	-	-	-	-
	QAE_{hyb}, $\alpha = 0.15,\ \beta = 1.5$	**74.3**	**78.2**	**-**	**-**	**-**	**-**
bge-large-en	-	61.9	66.4	-	-	-	-
	QAE_{hyb}, $\alpha = 0.3,\ \beta = 0.75$	**71.3**	**75.4**	**-**	**-**	**-**	**-**
bge-base-zh-v1.5	-	-	-	-	-	43.4	44.9
	QAE_{hyb}, $\alpha = 0.6,\ \beta = 1.0$	**-**	**-**	**-**	**-**	**48.6**	**49.5**
bge-base-zh	-	-	-	-	-	40.3	41.9
	QAE_{hyb}, $\alpha = 0.6,\ \beta = 1.5$	**-**	**-**	**-**	**-**	**48.5**	**49.5**
bge-base-en-v1.5	-	66.4	70.5	-	-	-	-
	QAE_{txt}, $\beta = 1.0$	**74.1**	**77.5**	-	-	**-**	**-**
bge-base-en	-	64.9	68.9	-	-	-	-
	QAE_{hyb}, $\alpha = 0.3,\ \beta = 1.5$	**71.7**	**75.9**	**-**	**-**	**-**	**-**
bce-embedding-base-v1	-	59.1	63.8	-	-	42.0	44.0
	QAE_{hyb}, $\alpha = 0.3,\ \beta = 0.5$	**66.8**	**71.1**	**-**	**-**	**47.0**	**48.3**
retriever-multi-base-256	-	33.8	37.4	-	-	-	-
	QAE_{emb}, $\alpha = 0.75$	**58.4**	**62.3**	**-**	**-**	**-**	**-**
quora-distilbert-base	-	29.8	33.2	-	-	-	-
	QAE_{emb}, $\alpha = 0.75$	**52.8**	**57.1**	**-**	**-**	**-**	**-**

Table 11: Retrieval performance of **monolingual and bilingual embedding models** on the latest datasets FIGNEWS(English) and CRUD-RAG(Chinese). Higher is better, with the best one bolded. '-' denotes default or null values due to mismatch between the language and model.