

Hyperspherical Semantic Expansion via von Mises–Fisher Mixtures for Zero-Shot Dense Retrieval

Sehong Oh

Graduate School of
Artificial Intelligence
Chung-Ang University
sehong@cau.ac.kr

Donghui Kim

Department of
Applied Data Science
Sungkyunkwan University
rookie0806@eg.skku.edu

Junseok Kwon

School of
Computer Science and
Engineering
Chung-Ang University
jskwon@cau.ac.kr

Abstract

Dense retrieval relies on aligning document and query embeddings; however, its performance is often constrained by the scarcity of labeled queries. Although prior work has attempted to model the many-to-one relationship between documents and queries using Gaussian Mixture Models (GMMs), such Euclidean-based approaches induce norm shrinkage ($\|\mu\| < 1$) in normalized semantic spaces, resulting in distorted similarity scores. To address this fundamental geometric inconsistency, we propose **vMF-HSE** (von Mises-Fisher for Hyperspherical Semantic Expansion), a training-free hyperspherical semantic expansion framework that models the latent query space using a von Mises-Fisher (vMF) mixture distribution on the unit hypersphere $\mathbb{S}^{d-1}$. We leverage Sra’s Newton-based optimization within the Expectation-Maximization (EM) algorithm to ensure robust parameter estimation even in high-dimensional semantic spaces. Building upon this stable estimation, we propose a novel adaptive stopping mechanism (κ -Stop) that adaptively determines the optimal number of synthetic queries per document by monitoring the convergence of marginal information gain. Experimental results on the BEIR benchmark demonstrate that vMF-HSE consistently outperforms conventional Euclidean-based methods in zero-shot retrieval settings, demonstrating the advantages of geometrically consistent modeling for normalized semantic spaces.

1 Introduction

Dense retrieval embeds queries and documents into a shared vector space, computing relevance through inner products (Karpukhin et al., 2020). This framework is foundational for downstream applications such as Retrieval-Augmented Generation (RAG) and Question Answering (Lewis et al., 2020). However, its effectiveness relies on annotated training

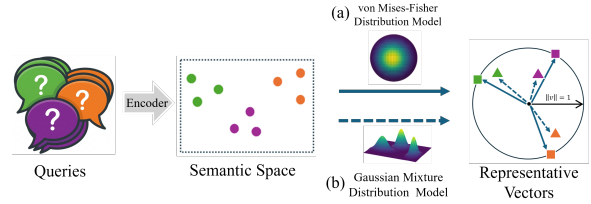


Figure 1: **Comparison of semantic expansion modeling on the unit hypersphere.** Traditional GMMs (triangles) exhibit norm shrinkage ($\|\mu\| < 1$) due to Euclidean averaging. In contrast, vMF-HSE (squares) maintains geometric consistency by modeling density directly on the hypersphere, ensuring $\|\mu\| = 1$.

data, which is often scarce. Furthermore, a structural mismatch exists between the semantic multiplicity of documents and the point-wise supervision typically used in training. While a single document often relates to a diverse set of queries (a many-to-one relationship), limited training pairs capture only a narrow slice of this semantic space. Consequently, aligning a document with a single query representation risks collapsing its semantic breadth, leading to biased retrieval signals.

Recent efforts have addressed this by modeling potential queries as probability distributions to capture multiple semantic facets. For instance, Potential Query Retrieval (PQR) represents the latent query space using Gaussian Mixture Models (GMMs) (Kang et al., 2025). While PQR enables a training-free, zero-shot framework, its reliance on Euclidean-based modeling introduces a geometric inconsistency with modern embedding spaces.

Modern embedding models generate L_2 -normalized vectors, where semantic information is primarily encoded in the direction (Xiao et al., 2024; Wang et al., 2022; Wang and Isola, 2020; Meng et al., 2019). In such spaces, the arithmetic mean employed by GMMs induces norm shrinkage ($\|\mu\| < 1$), causing representative vectors of semantically diverse clusters to drift toward the ori-

gin. As shown in Figure 1, this shrinkage occurs because Euclidean averaging disregards the intrinsic spherical geometry, resulting in mean vectors that deviate from the unit hypersphere. Thus, semantic intent (direction) becomes entangled with distributional variance (magnitude). Since the magnitude of an averaged vector scales inversely with its semantic dispersion, dense retrievers misinterpret this variance-driven shrinkage as a lack of relevance, penalizing documents with broader semantic coverage (Mardia and Jupp, 2000).

To address these limitations, we propose **vMF-HSE** (von Mises-Fisher for Hyperspherical Semantic Expansion), a training-free framework that models potential queries using a von Mises-Fisher mixture distribution on the unit hypersphere ($\mathbb{S}^{d-1}$) (Banerjee et al., 2005). By preserving manifold geometry, vMF-HSE decouples semantic intent (μ) from distributional concentration (κ). We utilize an Expectation-Maximization (EM) algorithm integrated with Sra’s Newton-based optimization (Banerjee et al., 2005) for stable κ estimation in high-dimensional settings. Additionally, we introduce the **κ -stop algorithm**, a dynamic stopping mechanism that adaptively terminates query expansion based on marginal information gain convergence, optimizing indexing efficiency without compromising retrieval performance.

Our contributions are as follows:

- **vMF-HSE Framework:** We introduce a training-free retrieval approach that employs vMF mixture distributions to eliminate norm shrinkage and preserve semantic integrity on the hypersphere.
- **Stable Parameter Estimation:** We incorporate Sra’s Newton-based optimization into the EM procedure, enabling robust estimation of the concentration parameter (κ) in high-dimensional spaces.
- **Adaptive Expansion Control (κ -stop):** We propose the κ -stop algorithm, which determines the optimal number of synthetic queries per document using marginal information gain convergence to balance retrieval quality and computational cost.
- **Empirical Validation:** Experimental results across multiple datasets show that vMF-HSE yields superior performance compared to GMM-based baselines in zero-shot retrieval (Thakur et al., 2021).

2 Related Work

Dense Retrieval. Dense retrieval has shifted the information retrieval paradigm from lexical matching, such as BM25 (Robertson and Walker, 1994), to semantic alignment in latent vector spaces (Karpukhin et al., 2020; Lee et al., 2019). Modern embedding models encode documents and queries into high-dimensional vectors, calculating similarity via inner products or cosine similarity (Reimers and Gurevych, 2019). These models are central to RAG systems, providing the context necessary for large language models (LLMs) to generate grounded responses (Lewis et al., 2020). However, traditional dense retrieval typically relies on point-to-point matching, which assumes a one-to-one relationship between a query and a document, often failing to capture the semantic breadth of complex documents.

Synthetic Query Generation and Diversity Optimization. To address training data sparsity and semantic multiplicity, LLMs have been utilized to generate synthetic queries. Methods like HyDE improve the alignment between user intent and document content by synthesizing pseudo-queries (Gao et al., 2022). Frameworks such as PQR leverage GMMs to model these queries and represent multiple semantic facets without additional training (Kang et al., 2025).

A key challenge remains in balancing relevance with diversity. While traditional approaches like Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) use explicit trade-offs, the recent Dartboard algorithm (Pickett et al., 2024) achieves diversity by optimizing for "relevant information gain" rather than using an explicit penalty. Unlike Dartboard, which focuses on result selection during the retrieval phase, vMF-HSE applies similar principles to the indexing phase to dynamically control synthetic query expansion.

Directional Modeling and Parameter Estimation. Estimating the parameters of vMF mixture models, particularly the mean direction μ and the concentration κ , presents significant computational challenges in high-dimensional spaces due to the involvement of modified Bessel functions (Banerjee et al., 2005). While the Expectation-Maximization (EM) algorithm is the standard framework for this task (Dhillon and Modha, 2001), numerical stability in estimating κ remains a critical issue. To address this, several robust optimization frameworks have been developed. Hornik and Grün (Hornik

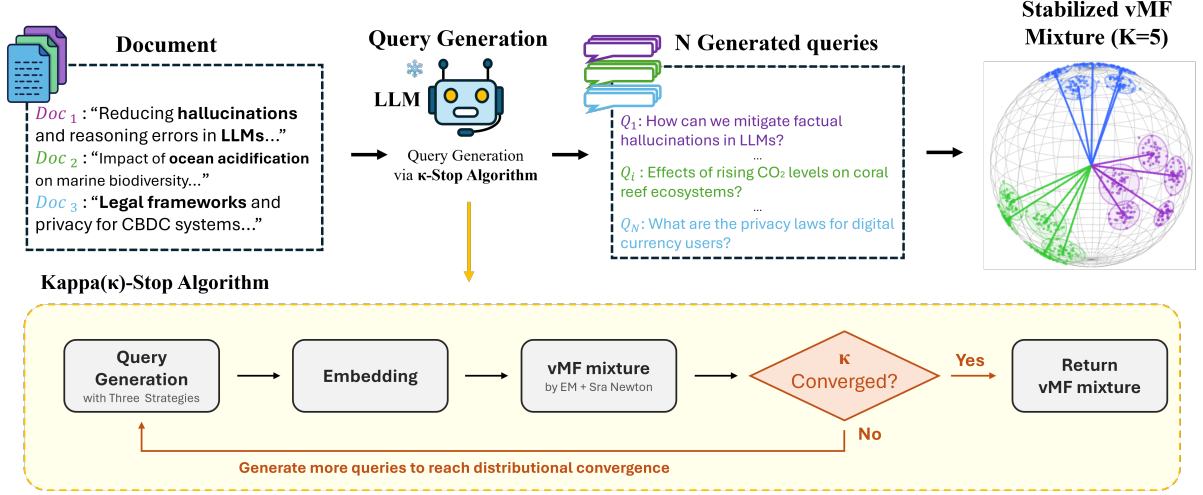


Figure 2: **Overview of the vMF-HSE framework.** The system samples potential queries from an input document using three distinct strategies—Zero-shot Sampling, Sliding-window Sampling, and Topic-aware Sampling—and maps them onto the unit hypersphere $\mathbb{S}^{d-1}$ as embeddings. The latent query space is iteratively modeled via a vMF mixture model, where parameters are estimated using an EM algorithm integrated with Sra-Newton optimization. The κ -Stop mechanism adaptively determines the expansion budget by monitoring the distributional convergence of the concentration parameter κ . Once stabilized, the framework returns the final vMF mixture, providing a robust and multi-faceted semantic representation of the document for retrieval tasks.

and Grün, 2014) introduced the movMF package, which has become a benchmark for stable EM-based fitting of vMF mixtures. Building on these numerical foundations, Gopal and Yang (Gopal and Yang, 2014) demonstrated the superiority of vMF-based clustering over Euclidean methods specifically for high-dimensional text data, highlighting its ability to capture latent semantic structures. Our framework utilizes the Newton-based approximation proposed by Sra (Sra, 2012) to achieve efficient and stable κ estimation. Unlike existing synthetic query expansion methods that rely on a fixed sampling budget (Kang et al., 2025), we leverage the convergence of these estimated parameters to implement an adaptive stopping criterion. Inspired by the principle of maximizing relevant information gain (Pickett et al., 2024), our κ -Stop algorithm terminates the expansion process once the distribution stabilizes, significantly optimizing indexing efficiency without sacrificing retrieval quality.

3 Methodology

Figure 2 illustrates the overall pipeline of the proposed method.

3.1 Synthetic Query Generation

To ensure a fair comparison with existing Gaussian-based frameworks, we adopt the same query generation strategies to construct the initial sample set $\mathcal{E}^{(0)}$. Specifically, we employ an LLM to generate

potential queries from a document D using three complementary strategies (Nogueira et al., 2019; Ma et al., 2021; Gao et al., 2022):

Zero-shot Sampling. The LLM is prompted to generate potential queries directly from the document. By exploiting stochasticity in the decoding process, we conduct N independent sampling trials to obtain a diverse set of query candidates (Gao et al., 2022).

Sliding-window Sampling. To capture semantic information at multiple granularities, we partition the document into coarse-, medium-, and fine-grained segments. This progressive refinement strategy (moving from global context to local details) is inspired by elaborative and rehearsal strategies in human cognitive psychology, which have been shown to enhance document representation and memorization in retrieval tasks (Tang et al., 2023; Lewis et al., 2020). By applying zero-shot sampling to each segment top-down, we ensure that the generated queries capture both the overarching theme and nuanced details of the document.

Topic-aware Sampling. The LLM first extracts key topics $T = \{t_1, t_2, \dots, t_k\}$ from the document. Conditioned on each topic and corresponding textual context, the model generates targeted queries, leveraging its reasoning capability to enhance both relevance and diversity (Ma et al., 2021).

By employing these strategies, we ensure consistency with the experimental protocols of baseline

models such as PQR (Kang et al., 2025). This alignment guarantees that any observed performance improvements stem from our hyperspherical modeling framework, rather than variations in the input data or query generation process.

3.2 Distribution Estimation via vMF-Sra

We assume that the potential queries associated with a document D are generated from a vMF mixture distribution rather than a Gaussian mixture. This modeling choice is motivated by the directional nature of L_2 -normalized embeddings, where semantic information is encoded on the unit hypersphere. In such spaces, Euclidean modeling disregards the manifold’s curvature, leading to information distortion through norm shrinkage.

The von Mises-Fisher Distribution. For a d -dimensional semantic space, the vMF distribution over a unit vector $\mathbf{x} \in \mathbb{S}^{d-1}$ is defined as:

$$f(\mathbf{x}; \boldsymbol{\mu}, \kappa) = C_d(\kappa) \exp(\kappa \boldsymbol{\mu}^T \mathbf{x}), \quad (1)$$

where $\boldsymbol{\mu}$ is the mean direction ($\|\boldsymbol{\mu}\| = 1$), $\kappa \geq 0$ is the concentration parameter, and $C_d(\kappa)$ is the normalization constant:

$$C_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)}, \quad (2)$$

where $I_v(\cdot)$ denotes the modified Bessel function of the first kind at order v . The overall distribution is modeled as a weighted mixture of K vMF components:

$$P(\mathbf{x}) = \sum_{k=1}^K \pi_k f(\mathbf{x}; \boldsymbol{\mu}_k, \kappa_k).$$

Parameter Estimation via EM. To estimate the parameters $\{\pi_k, \boldsymbol{\mu}_k, \kappa_k\}$, we employ the EM algorithm as follows:

- **E-Step:** We compute the posterior responsibility γ_{ik} , representing the probability that the k -th component generates the i -th sample $\mathbf{x}_i$:

$$\gamma_{ik} = \frac{\pi_k f(\mathbf{x}_i; \boldsymbol{\mu}_k, \kappa_k)}{\sum_{j=1}^K \pi_j f(\mathbf{x}_i; \boldsymbol{\mu}_j, \kappa_j)}. \quad (3)$$

For numerical stability, this is performed in the log-likelihood domain.

- **M-Step:** We update the parameters by maximizing the expected log-likelihood. The weights π_k are updated as N_k/M , while the

mean direction $\boldsymbol{\mu}_k$ is obtained by projecting the weighted resultant vector $\bar{\mathbf{r}}_k = \sum_i \gamma_{ik} \mathbf{x}_i$ onto the unit hypersphere:

$$\boldsymbol{\mu}_k = \frac{\bar{\mathbf{r}}_k}{\|\bar{\mathbf{r}}_k\|}. \quad (4)$$

Decoupling Direction and Concentration. A fundamental limitation of Euclidean mixture modeling is norm shrinkage. In GMMs, the representative vector $\mathbf{v}_k$ is calculated as a weighted arithmetic mean. For unit vectors $\mathbf{x}_i$, Jensen’s inequality implies that the norm of their (non-identical) mean is strictly less than one:

$$\|\mathbf{v}_k^{\text{GMM}}\| = \left\| \frac{1}{N_k} \sum \gamma_{ik} \mathbf{x}_i \right\| < 1. \quad (5)$$

This shrinkage entangles semantic direction with dispersion, since reduced norm implicitly reflects variance. As a result, similarity scores become distorted: magnitude attenuation is incorrectly interpreted as directional change (Wang and Isola, 2020).

In contrast, vMF-HSE explicitly decouples direction and concentration by estimating the concentration parameter κ_k separately from the mean direction $\boldsymbol{\mu}_k$ using the mean resultant length $\bar{R}_k = \frac{\|\bar{\mathbf{r}}_k\|}{N_k}$ (Mardia and Jupp, 2000; Banerjee et al., 2005). Specifically, κ_k is obtained by solving

$$A_d(\kappa) = \bar{R}_k, \quad (6)$$

where $A_d(\kappa) = \frac{I_{d/2}(\kappa)}{I_{d/2-1}(\kappa)}$. Since this equation does not admit a closed-form solution, we employ Sra’s Newton-based iterative method to solve for κ (Sra, 2012):

$$\kappa^{(m+1)} = \kappa^{(m)} - \frac{A_d(\kappa^{(m)}) - \bar{R}_k}{A_d'(\kappa^{(m)})}. \quad (7)$$

Here, $A_d'(\kappa) = 1 - A_d(\kappa)^2 - \frac{d-1}{\kappa} A_d(\kappa)$ and m denotes the Newton iteration index. This formulation guarantees that each mean direction $\boldsymbol{\mu}_k$ remains on the unit hypersphere, while dispersion is captured exclusively by κ_k . Consequently, semantic direction is preserved without the magnitude attenuation inherent in Euclidean GMMs.

3.3 Concentration-Based Adaptive Stopping (κ -Stop)

To improve efficiency, we introduce κ -Stop, an adaptive stopping criterion leveraging the stability

Algorithm 1 Concentration-based Adaptive Stopping (κ -Stop)

Require: Document d , Components K , Patience P , Threshold τ , Batch size s , Initial sample size n_0 , Strategy set $\mathcal{S}$

Ensure: Optimized vMF mixture centers $\{\mu_k\}_{k=1}^K$

- 1: $\mathcal{E}^{(0)} \leftarrow \bigcup_{strat \in \mathcal{S}} \{E(G_{strat}(d, \lfloor n_0/|\mathcal{S}| \rfloor))\} \triangleright$
Initialize with multi-granular coverage
- 2: Fit vMF on $\mathcal{E}^{(0)} \rightarrow$ calculate $\bar{\kappa}^{(0)}$
- 3: $c \leftarrow 0, \quad t \leftarrow 1$
- 4: **repeat**
- 5: $\mathcal{Q}_{new} \leftarrow \bigcup_{strat \in \mathcal{S}} G_{strat}(d, \lfloor s/|\mathcal{S}| \rfloor) \triangleright$
Generate stratified batch across strategies
- 6: $\mathcal{E}^{(t)} \leftarrow \mathcal{E}^{(t-1)} \cup \{E(q) \mid q \in \mathcal{Q}_{new}\}$
- 7: Update vMF on $\mathcal{E}^{(t)} \rightarrow$ obtain $\{\kappa_k^{(t)}\}_{k=1}^K$
- 8: $\bar{\kappa}^{(t)} \leftarrow \frac{1}{K} \sum_{k=1}^K \kappa_k^{(t)}$
- 9: $\delta^{(t)} \leftarrow |\bar{\kappa}^{(t)} - \bar{\kappa}^{(t-1)}| / \bar{\kappa}^{(t-1)}$
- 10: **if** $\delta^{(t)} < \tau$ **then**
- 11: $c \leftarrow c + 1$
- 12: **else**
- 13: $c \leftarrow 0$
- 14: **end if**
- 15: $t \leftarrow t + 1$
- 16: **until** $c \geq P$ **or** maximum budget reached
- 17: **return** $\{\mu_k\}_{k=1}^K$

of the concentration parameter κ . We quantify semantic saturation by tracking the relative change in average concentration:

$$\delta^{(t)} = \frac{|\bar{\kappa}^{(t)} - \bar{\kappa}^{(t-1)}|}{\bar{\kappa}^{(t-1)}}, \quad \text{where} \quad \bar{\kappa}^{(t)} = \frac{1}{K} \sum_{k=1}^K \kappa_k^{(t)}. \quad (8)$$

As detailed in Algorithm 1, termination is triggered when $\delta^{(t)} < \tau$ for P consecutive steps. This mechanism allocates resources adaptively based on document complexity, preserving retrieval quality while significantly optimizing indexing costs.

4 Experiments

4.1 Experimental Setup

We evaluate vMF-HSE on five datasets from the BEIR benchmark (Thakur et al., 2021): FiQA, ArguAna, SciDocs, SciFact, and NFCorpus, using NDCG@10 as the primary metric. For each

Table 1: **GMM norm shrinkage statistics** ($K=5$). *Shrink*: $1 - \|\mu\|$. *Trace*: $\text{Tr}(\Sigma)$. ρ : Correlation between $\|\mu\|$ and $\text{Tr}(\Sigma)$.

Dataset	Encoder	$\ \mu\ $	Std	Shrink	Trace	ρ
SciFact	BGE-M3	0.814	0.054	18.6%	0.336	-0.999
	MiniLM	0.816	0.076	18.4%	0.329	-0.997
	Sent-T5	0.939	0.017	6.1%	0.118	-1.000
NFCorpus	BGE-M3	0.831	0.055	16.9%	0.307	-0.999
	MiniLM	0.836	0.071	16.4%	0.296	-0.998
	Sent-T5	0.946	0.016	5.5%	0.107	-1.000
ArguAna	BGE-M3	0.801	0.049	19.9%	0.357	-0.999
	MiniLM	0.791	0.071	21.0%	0.371	-0.998
	Sent-T5	0.934	0.016	6.6%	0.128	-1.000
FiQA	BGE-M3	0.810	0.057	19.0%	0.341	-0.998
	MiniLM	0.771	0.068	22.9%	0.402	-0.997
	Sent-T5	0.935	0.017	6.5%	0.127	-1.000
SciDocs	BGE-M3	0.809	0.055	19.1%	0.344	-0.999
	MiniLM	0.785	0.088	21.5%	0.376	-0.997
	Sent-T5	0.935	0.017	6.5%	0.127	-1.000

document, we generate 300 synthetic queries using Llama-3.1-8B-Instruct ($T = 1.2$) through zero-shot, sliding-window, and topic-aware sampling. These queries are encoded using three models with varying capacities: Sentence-T5-base (Ni et al., 2021), BGE-M3 (Chen et al., 2024), and all-MiniLM-L12-v2 (Reimers and Gurevych, 2019).

We compare our method against several baseline categories: traditional lexical retrieval (BM25), zero-shot generative expansion (docT5query, HyDE), supervised neural retrievers (SPLADE, ANCE, RocketQAv2), and late-interaction models (ColBERTv2, CITADEL). We also include PQR (Kang et al., 2025), the most closely related framework utilizing Gaussian mixtures. To isolate the impact of distribution modeling, we compare the *Avg. Base Encoder*, which directly encodes the original document text, against expansion-based variants: a standard Euclidean *GMM* and a *GMM-N* variant where centroids are projected onto the unit hypersphere ($\mu_k \leftarrow \mu_k / \|\mu_k\|$).

While GMM-N serves as a heuristic correction for norm shrinkage, it remains inconsistent with Euclidean maximum likelihood estimation. In contrast, vMF-HSE models the query distribution directly on the hypersphere, ensuring that both parameter estimation and representation remain geometrically consistent with cosine-based retrieval. All methods are evaluated in a zero-shot setting without task-specific fine-tuning. Unless otherwise specified, we use $K = 5$ mixture components for all distribution-based methods. All experiments were conducted on a single NVIDIA A100 GPU.

Table 2: **Zero-shot retrieval results on the BEIR benchmark** ($K=5$, 300 queries per document, NDCG@10). Bold and underline mark the best and second-best averages among Base Encoder, GMM, GMM-N, and vMF.

Model	FiQA	ArguAna	SciDocs	SciFact	NFCorpus	Avg.
BM25	0.308	0.285	0.312	0.664	0.540	0.422
docT5query	0.293	0.351	0.160	0.678	0.330	0.362
SPLADE	0.292	0.448	0.151	0.635	0.325	0.370
ANCE	0.368	0.325	0.284	0.529	0.430	0.387
RocketQAv2	0.305	0.454	0.134	0.570	0.295	0.352
ColBERTv2	0.358	0.465	0.157	0.696	0.340	0.403
CITADEL	0.335	0.493	0.150	0.698	0.339	0.403
HyDE _{Llama-8B}	0.225	0.305	0.138	0.645	0.318	0.326
Sentence-T5	0.460	0.319	0.107	0.426	0.481	0.359
BGE-M3	0.531	0.382	0.128	0.635	0.523	0.440
MiniLM	0.492	0.359	0.159	0.631	0.525	0.433
<i>Avg. Base Encoder</i>	<u>0.494</u>	<u>0.353</u>	<u>0.131</u>	<u>0.564</u>	<u>0.510</u>	<u>0.411</u>
GMM _{Sentence-T5}	0.436	0.361	0.141	0.465	0.525	0.386
GMM _{BGE-M3}	0.441	0.385	0.158	0.649	0.547	0.436
GMM _{MiniLM}	0.452	0.384	0.198	0.668	0.563	0.453
<i>Avg. GMM</i>	<u>0.443</u>	<u>0.377</u>	<u>0.166</u>	<u>0.594</u>	<u>0.545</u>	<u>0.425</u>
GMM-N _{Sentence-T5}	0.445	0.375	0.140	0.462	0.510	0.386
GMM-N _{BGE-M3}	0.455	0.401	0.165	0.647	0.535	0.441
GMM-N _{MiniLM}	0.465	0.405	0.195	0.665	0.548	0.456
<i>Avg. GMM-N</i>	<u>0.455</u>	<u>0.394</u>	<u>0.167</u>	<u>0.591</u>	<u>0.531</u>	<u>0.428</u>
vMF _{Sentence-T5}	0.498	0.412	0.158	0.513	0.510	0.418
vMF _{BGE-M3}	0.515	0.428	0.176	0.655	0.567	0.468
vMF _{MiniLM}	0.517	0.412	0.210	0.660	0.562	0.472
<i>Avg. vMF</i>	0.510	0.417	0.181	0.609	0.546	0.453

4.2 Norm Shrinkage Analysis

For data points on the unit hypersphere $\mathcal{S}^{d-1}$, there exists a fundamental coupling between the sample mean and covariance of each GMM component. Given N unit-norm vectors $\{\mathbf{x}_i\}_{i=1}^N$ assigned to a component with mean $\boldsymbol{\mu} = \frac{1}{N} \sum_i \mathbf{x}_i$, the trace of the sample covariance satisfies the following identity: $\text{Tr}(\boldsymbol{\Sigma}) = 1 - \|\boldsymbol{\mu}\|^2$. This identity reveals that GMM inherently *couples* direction and concentration: as the within-cluster variance increases ($\text{Tr}(\boldsymbol{\Sigma}) \uparrow$), the norm of the mean necessarily decreases ($\|\boldsymbol{\mu}\| \downarrow$). Such norm shrinkage distorts similarity scores in dense retrieval, as inner products become entangled with distributional dispersion. In contrast, vMF preserves geometric consistency by constraining all component means to lie on the unit hypersphere ($\|\boldsymbol{\mu}_k\| \equiv 1$), thereby effectively *decoupling* semantic intent from concentration through the independent parameter κ_k .

Table 1 empirically validates this phenomenon. Across all evaluated datasets, GMM components exhibit substantial norm shrinkage (up to 22.9%), while the near-perfect negative correlation ($\rho \leq -0.997$) confirms the deterministic coupling between variance and magnitude.

4.3 Retrieval Performance

Table 2 compares the proposed vMF-HSE with established retrieval baselines. Across five zero-shot retrieval datasets, vMF-HSE achieves an average

NDCG@10 of 0.453, consistently outperforming both the base encoders (0.411) and GMM-based expansion models (0.426).

To isolate the impact of directional modeling from the norm shrinkage effect, we evaluated GMM-N, constructed by applying L_2 normalization to the pseudo-query vectors generated by standard GMM expansion. While this post-hoc normalization yields marginal improvements on certain datasets, such as ArguAna, it fails to produce meaningful gains in the overall average (0.428). These results suggest that simply projecting Euclidean-based representations onto a hypersphere is insufficient to resolve geometric inconsistencies.

In contrast, vMF-HSE yields substantial and consistent improvements by inherently preserving directional consistency throughout the density estimation process. Notably, all variants of vMF-HSE outperform their GMM and GMM-N counterparts across all backbone architectures. This demonstrates that hyperspherical modeling provides a more effective and robust framework for generative document expansion in dense retrieval.

To verify that these gains are statistically stable, we repeat the main experiments over 10 independent runs with different random seeds for the LLM-based query generation pipeline. The resulting standard deviations are consistently small (± 0.007 – 0.010 in absolute NDCG@10; Table 4), and the relative ordering among all configurations

Table 3: **Performance of κ -based adaptive stopping under different convergence thresholds (τ) with patience $P=3$.** Parenthesized percentages indicate NDCG@10 relative to the best Fixed- N (vs. Best). Saving: percentage of queries not generated due to early stopping.

		ArguAna		SciFact		NFCorpus	
		NDCG@10	Saving	NDCG@10	Saving	NDCG@10	Saving
<i>Best Fixed-N</i> ($N=10$ –250)		0.4096	—	0.6645	—	0.5565	—
τ	0.01	0.4133(100.9%)	32.8%	0.6625(99.7%)	45.8%	0.5543(99.6%)	60.9%
	0.02	0.4100(100.1%)	57.4%	0.6585(99.1%)	70.5%	0.5586(100.4%)	77.2%
	0.05	0.3940(96.2%)	84.8%	0.6572(98.9%)	86.4%	0.5459(98.1%)	88.1%
	0.10	0.3883(94.8%)	91.9%	0.6353(95.6%)	89.8%	0.5454(98.0%)	90.5%

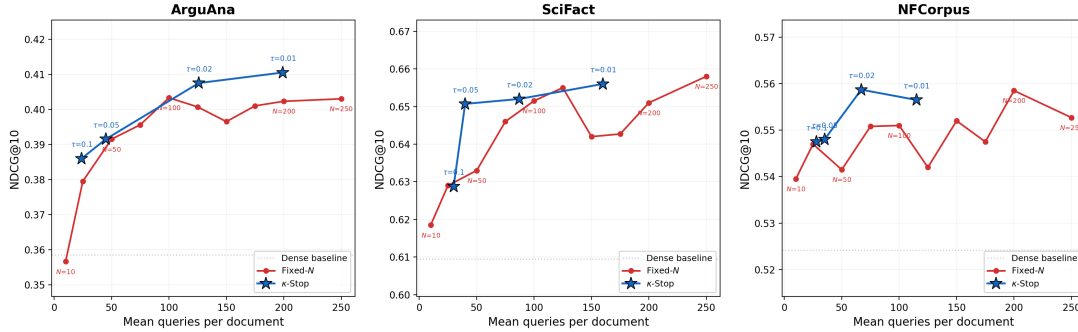


Figure 3: **Comparison of Fixed- N and the proposed κ -Stop in retrieval performance versus mean queries per document** across three BEIR datasets. While Fixed- N uses a uniform budget, κ -Stop adaptively terminates query allocation when the vMF posterior convergence reaches a predefined threshold τ , achieving efficiency.

is preserved across runs, confirming that the reported improvements are not driven by favorable seeds. Furthermore, to assess generalization beyond the five datasets, we additionally evaluate on two CQADupStack sub-benchmarks — Webmasters ($\sim 17K$ documents) and English ($\sim 40K$ documents) — whose domains differ substantially from the scientific, financial, and argumentative settings of the original five. vMF-HSE achieves the best average NDCG@10 on both benchmarks (0.4288 and 0.4982, respectively), outperforming GMM and GMM- N ; full results are provided in Appendix E.

4.4 Comparison with Naive Expansion

To isolate the contribution of vMF mixture modeling from the general benefit of having more synthetic queries, we compare vMF-HSE ($K=5$) against a *Naive Expansion* baseline that uses all $N=300$ raw synthetic queries directly for retrieval, without clustering or refinement. Table 4 reports results over 10 independent runs.

The comparison reveals three points. First, hyperspherical modeling contributes beyond expansion alone: the Naive baseline captures the general benefit of synthetic query expansion (+0.021 over the base encoder; 0.432 vs. 0.411 in Table 2), but vMF-HSE adds a further +0.023 on top of that (0.455 vs. 0.432), roughly doubling the gain — the

geometric reformulation contributes its own distinct improvement rather than merely benefiting from more queries. The one exception is SciFact (0.627 vs. 0.614), where the concentrated semantic content reduces the benefit of clustering; vMF-HSE is consistently better on the other four datasets.

Second, vMF-HSE acts as a denoising filter. Using all 300 raw queries introduces redundancy and stochastic noise from the LLM: by distilling them into K directional centroids on the unit hypersphere, vMF-HSE reduces run-to-run variance by roughly 30% ($\pm 0.013 \rightarrow \pm 0.009$), stabilizing the document representation. Third, Naive retrieval requires comparing the user query against all 300 vectors per document, which does not scale to large corpora; vMF-HSE achieves comparable or better quality with only $K=5$ centroids — a $60\times$ reduction in per-document comparisons — so the quality gain incurs no additional retrieval-time cost.

4.5 Non-Normalized Retrievers

Since our motivation relies on the directional nature of L_2 -normalized embedding spaces, we examine whether the advantage of vMF-HSE persists on DPR (Karpukhin et al., 2020), a representative dense retriever whose embeddings are not explicitly normalized, using three representative datasets (ArguAna, NFCorpus, and SciFact).

Table 4: **Naive Expansion** ($N=300$) vs. **vMF-HSE** ($K=5$) — **NDCG@10**. Mean $\pm$ std over 10 seeds, averaged across three encoders. Per-encoder results in Appendix D.

Method	FiQA	ArguAna	SciDocs	SciFact	NFCorpus	Avg.
Base Enc.	0.494	0.353	0.131	0.564	0.510	0.411
Naive	0.476 $\pm$.013	0.391 $\pm$.010	0.135 $\pm$.016	0.627 $\pm$.012	0.532 $\pm$.013	0.432 $\pm$.013
vMF (Ours)	0.511 $\pm$.008	0.421 $\pm$.007	0.185 $\pm$.009	0.614 $\pm$.009	0.550 $\pm$.010	0.455 $\pm$.009

Table 5: **Applicability to a non-normalized retriever (DPR)** — **NDCG@10**. All variants are evaluated under the same setting as Table 2 ($K=5$, $N=300$).

Method	ArguAna	NFCorpus	SciFact	Avg.
Baseline	0.1835	0.3571	0.1829	0.2412
GMM _{raw}	0.2511	0.3759	0.3105	0.3125
GMM _{normalized}	0.2556	0.3730	0.2994	0.3093
vMF (Ours)	0.2619	0.3851	0.3028	0.3166

As shown in Table 5, vMF-HSE achieves the highest average NDCG@10 (0.3166), outperforming both raw GMM (0.3125) and normalized GMM (0.3093), with all mixture-based variants improving over the DPR baseline by 28–66%. The only exception is SciFact, where vMF trails GMM_{raw} by 0.008 — within the variance observed in Table 4. The reason is mechanistic: the vMF M-step projects each mean direction onto the unit hypersphere (Eq. 4), so the norm shrinkage documented in Table 1 is structurally avoided regardless of whether the input embeddings are normalized. The unit-hypersphere assumption is thus not a precondition for our method — it suffices that the latent query space be modeled on the hypersphere, which our framework guarantees by construction.

4.6 Efficiency of the κ -Stop Mechanism

We evaluate the efficiency of the κ -stop mechanism by analyzing the trade-off between retrieval performance and the number of synthetic queries. As illustrated in Figure 3, while a fixed-budget strategy (Fixed- N) requires a uniform and often excessive number of queries to reach peak performance, κ -stop adaptively terminates expansion based on vMF posterior convergence. This allows vMF-HSE to achieve performance comparable to or better than the full query budget across all datasets while using significantly fewer samples.

Table 3 provides a quantitative breakdown of this efficiency under different convergence thresholds (τ). At $\tau = 0.01$, κ -stop retains nearly 100% of maximum accuracy while reducing query counts by 32.8% for ArguAna and 60.9% for NFCorpus. Increasing the threshold to $\tau = 0.02$ further improves efficiency, yielding a 77.2% reduction in NFCorpus while maintaining 100.4%

relative NDCG@10 compared to the best Fixed- N baseline. These results indicate that κ -stop adapts to document characteristics, allocating more samples to topically diverse documents while converging quickly for concentrated ones, and thereby acts as an implicit noise filter. Beyond query savings, an end-to-end per-document indexing cost analysis and a sensitivity study of the patience parameter P are provided in Appendix F.

4.7 Impact of Cluster Cardinality (K)

Table 6 shows how retrieval performance varies as the number of clusters K increases from 3 to 8. We observe that vMF-HSE is robust to the choice of K ; for instance, NDCG@10 on NFCorpus remains stable within a narrow range (0.5545–0.5669). This stability indicates that our hyperspherical framework can effectively represent document semantics without extensive hyperparameter tuning.

For practical applications, we automate the selection of K to ensure scalability. Although the Bayesian Information Criterion (BIC) (Schwarz, 1978) is a standard metric, its classical penalty term, $\frac{M}{2} \ln N$, can be overly restrictive in high-dimensional settings. Since the number of free parameters M grows linearly with the dimension d , standard BIC tends to underestimate the optimal number of clusters, leading to suboptimal performance on complex datasets such as SciFact.

To address this, we employ k -log-BIC with a modified penalty term. By scaling the complexity cost logarithmically with K , k -log-BIC enables the model to capture a diverse set of semantic facets within a document. This leads to improved retrieval accuracy, as evidenced by NDCG@10 of 0.6696 on SciFact. These results suggest that while the vMF mixture is inherently stable, the k -log-BIC criterion plays a key role in capturing fine-grained sub-topics necessary for effective zero-shot retrieval.

4.8 Effect of Query Sampling Strategy

We investigate the impact of different query sampling strategies—zero-shot, sliding-window, and topic-aware—on retrieval performance. As shown in Table 7, the optimal strategy varies across datasets: topic-aware sampling excels on SciFact, whereas sliding-window is most effective for NFCorpus. Notably, the combined approach achieves the highest average performance (0.548), suggesting that integrating multiple generation perspectives provides more comprehensive coverage of the document’s latent query space.

Table 6: **Comparison of NDCG@10 between GMM and vMF across different K values** using the BGE-M3 encoder. BIC and k -log-BIC denote automated K selection methods.

	Method	SciFact		NFCorpus		ArguAna	
		GMM	vMF	GMM	vMF	GMM	vMF
Fixed K	$K = 3$	0.6458	0.6347	0.5419	0.5545	0.3827	0.4219
	$K = 4$	0.6421	0.6404	0.5565	0.5585	0.3797	0.4232
	$K = 5$	0.6458	0.6552	0.5570	0.5669	0.3854	0.4281
	$K = 6$	0.6409	0.6668	0.5562	0.5629	0.3861	0.4212
	$K = 7$	0.6405	0.6643	0.5555	0.5610	0.3848	0.4217
	$K = 8$	0.6498	0.6567	0.5564	0.5656	0.3864	0.4243
Auto	BIC	0.6599	0.6511	0.5465	0.5520	0.3829	0.4238
	k -log-BIC	0.6599	0.6696	0.5465	0.5671	0.3829	0.4222

Table 7: **Ablation study of query sampling strategies.** Results are reported in NDCG@10 using vMF-HSE with BGE-M3 ($K = 5$). The 'Combined' strategy integrates samples from all three methods.

Strategy	SciFact	NFCorpus	ArguAna	Avg
Zero-shot	0.658	0.555	0.419	0.544
Sliding	0.645	0.563	0.409	0.539
Topic	0.661	0.537	0.412	0.537
Combined	0.657	0.557	0.430	0.548

4.9 Impact of LLM Architecture and Scale

We analyze how the choice of LLM backbone and parameter scale affects retrieval performance. As shown in Figure 4(a), models in the 7–9B range exhibit comparable performance with minor domain-specific variations. Mistral-7B (0.6695) slightly outperforms others on SciFact, while Llama-3.1-8B achieves the best performance on NFCorpus (0.5569) and ArguAna (0.4225), likely due to its superior reasoning in argumentative structures.

The effect of model scale, evaluated using the Llama family (Figure 4b), shows a clear correlation between model size and effectiveness. Scaling down from 8B to 1B results in performance declines of 2.6%, 2.9%, and 3.6% on SciFact, NFCorpus, and ArguAna, respectively. Although the 1B model still improves upon the base encoder, its gain on challenging tasks such as ArguAna is marginal (+0.005). These results suggest that while vMF-HSE is architecture-agnostic, larger backbones produce more semantically informative queries.

5 Conclusion

In this paper, we introduced vMF-HSE, a training-free framework for hyperspherical semantic expansion in dense retrieval. By modeling the latent query space with von Mises–Fisher mixtures, our approach addresses norm shrinkage in Euclidean Gaussian models and preserves semantic integrity by decoupling direction from concentration. Experiments on the BEIR benchmark show that vMF-

vMF-HSE Retrieval Performance across Query Generators (BGE-M3)

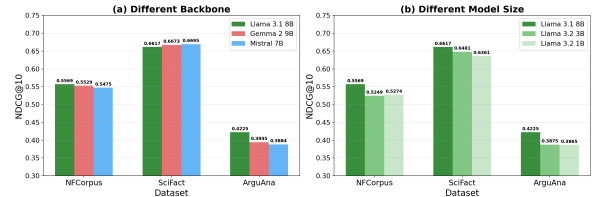


Figure 4: **Impact of LLM architecture and scale on retrieval performance.** (a) Comparison of different LLM backbones with similar parameter scales (7–9B). (b) Performance scaling across the Llama models (1B, 3B, and 8B).

HSE consistently outperforms base encoders and GMM-based methods across multiple architectures. Additionally, the proposed κ -Stop mechanism improves indexing efficiency by reducing query expansion with minimal performance loss. These results highlight the effectiveness of geometrically consistent modeling for scalable zero-shot retrieval.

6 Limitations

Although our framework yields strong empirical results, certain limitations remain. Specifically, while k -log-BIC alleviates over-penalization, it still relies on a heuristic that may require tuning, which potentially restricts full automation. Additionally, the performance of our approach remains dependent on the LLM backbone, which may introduce model-specific biases into the synthetic queries. Future work will explore non-parametric approaches for automated cluster discovery.

7 Acknowledgments

This work was supported in part by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2025-02217071), and in part by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2021-II211341, RS-2025-25422680).

References

- Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. 2005. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6:1345–1382.
- Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *ACM SIGIR*, pages 335–336.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of ACL*, pages 2318–2335.
- Inderjit S. Dhillon and Dharmendra S. Modha. 2001. Concept decompositions for large sparse text data using clustering. In *Machine Learning*, volume 42, pages 143–175.
- Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Precise zero-shot dense retrieval without relevance labels. *arXiv preprint arXiv:2212.10496*.
- Siddharth Gopal and Yiming Yang. 2014. von mises-fisher clustering models. In *ICML*, pages 154–162.
- Kurt Hornik and Bettina Grün. 2014. movmf: An r package for fitting mixtures of von mises-fisher distributions. *Journal of Statistical Software*, 58:1–31.
- Junfeng Kang, Rui Li, Qi Liu, Yanjiang Chen, Zheng Zhang, Junzhe Jiang, Heng Yu, and Yu Su. 2025. Pqr: Improving dense retrieval via potential query modeling. In *ACL*, pages 13455–13469.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP*, pages 6769–6781.
- Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *ACL*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*, pages 9459–9474.
- Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 2021. Zero-shot neural passage retrieval via domain-targeted synthetic question generation. In *EACL*, pages 1075–1088.
- Kanti V. Mardia and Peter E. Jupp. 2000. *Directional Statistics*. Wiley.
- Yu Meng, Jiaxin Huang, Guanganyuan Wang, Zihan Wang, Chao Zhang, Yu Zhang, and Jiawei Han. 2019. Spherical text embedding. In *NeurIPS*, pages 14993–15003.
- Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B. Hall, Daniel Cer, and Yinfei Yang. 2021. Sentence-T5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*.
- Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*.
- Marc Pickett, Jeremy Hartman, Ayan Kumar Bhowmick, Raquib-ul Alam, and Aditya Vempaty. 2024. Better rag using relevant information gain. *arXiv preprint arXiv:2407.12101*.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP*, pages 3982–3992.
- Stephen Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *ACM SIGIR*, pages 232–241.
- Gideon Schwarz. 1978. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.
- Suvrit Sra. 2012. A short note on parameter approximation for von mises-fisher distributions. *Computational Statistics*, 27(1):177–190.
- Yubao Tang, Ruqing Wang, Yixuan Liu, and 1 others. 2023. Semantic-enhanced differentiable search index inspired by learning strategies. In *ACM SIGKDD*.
- Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *NeurIPS Datasets and Benchmarks Track*.
- 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*.
- Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, pages 9929–9939.
- Shitao Xiao, Zheng Liu, Peitian Zhang, and Xingrun Hou. 2024. C-pack: Packed resources for general chinese embeddings. In *ACM SIGIR*, pages 5521–5532.

A Other Implementation Details

This section reports implementation details used in our vMF-HSE pipeline.

- **Generator and decoding.** We use meta-llama/Llama-3.1-8B-Instruct with temperature $T=1.2$, top_p= 0.9, and max_new_tokens= 28.
- **Sampling budget.** We use 100 samples per strategy (Zero-shot, Sliding-window, Topic-aware), yielding about 300 candidate queries per document before de-duplication.
- **Document truncation.** In the generation code, each input passage is clipped to the first 6000 characters (used as the practical implementation of the 6K-length cap).
- **Sliding-window sampling.** We split documents into sentences and use three granularities with steps {4, 2, 1} (fine, medium, coarse). Each window contains at least 5 sentences. We use half-window stride and cap fragments per level for stable throughput.
- **Topic-aware sampling.** We first sample 5 topic strings, then allocate queries uniformly across discovered topics.
- **vMF estimation (Sra-Newton).** We fit vMF mixtures with EM and Newton-based κ updates using max_iter= 100, tol= 10^{-4} , newton_iter= 10, kappa_max= 1000, and random_state= 42.
- **GMM baseline.** For Euclidean modeling, we use sklearn.mixture.GaussianMixture with covariance_type='diag', max_iter= 100, and random_state= 42.
- **Normalization.** All query/document embeddings and mixture centers are L2-normalized before similarity scoring.

B Example Display

We show one NFCorpus example document and sampled potential queries from each strategy.

Document ID: MED-1798

Adenovirus 36 infection and obesity. ... We reviewed the literature concerning the role of adenovirus 36 (AD-36), the most widely studied infectious agent in animals and humans, because of its potential association with childhood obesity. The available evidence suggests that more studies are needed to evaluate whether the association between AD-36 antibodies and obesity is causal or coincidental.

Zero-shot Sampling (examples)

1. Can experimental data support the relationship between viral infections and obesity?
2. What role may adenovirus 36 play in the development of childhood obesity?
3. Can infections, like adenovirus 36, cause weight gain in children?
4. What is known about the association between adenovirus 36 infection and obesity in children?

Sliding-window Sampling (examples)

1. Do adenovirus infections have a link to obesity in human subjects?
2. What do scientific studies suggest regarding the association between adenovirus 36 infection and human obesity?
3. Has adenovirus 36 been associated with obesity in human studies?
4. What evidence does the literature suggest about the association between adenovirus 36 antibodies and obesity?

Topic-aware Sampling (examples)

1. Is the presence of adenovirus 36 (AD-36) associated with obesity?
2. What is known about the relationship between Adenovirus 36 infection and obesity in children?
3. Is adenovirus 36 infection associated with a higher risk of obesity in children?
4. Can an adenovirus be a cause of obesity in children?

C Prompting Templates

Zero-shot and Sliding-window templates. The same dataset-specific prompt is used in both strategies (with different passage spans).

NFCorpus

```
{PASSAGE}
Please read the above passage about medical
information, and write a Question from a
different perspective that a dense retrieval
model could use to find this Passage. Directly
output the Question without any additional
information.
Question:
```

SciFact

```
{PASSAGE}
Please read the above Passage about a
scientific claim, and write a Question from a
different perspective that a dense retrieval
model could use to find this Passage. Directly
output the Question without any additional
information.
Question:
```

SciDocs

```
{PASSAGE}
Please read the above Passage about a
scientific claim, and write a Question from a
different perspective that a dense retrieval
model could use to find this Passage. Directly
output the Question without any additional
information.
Question:
```

ArguAna

```
{PASSAGE}
Please read the above Passage, and write a
Question from a different perspective that a
dense retrieval model could use to find this
Passage. Directly output the Question without
any additional information.
Question:
```

FiQA

```
{PASSAGE}
Please read the above Passage about financial
information, and write a Question from a
different perspective that a dense retrieval
model could use to find this Passage. Directly
output the Question without any additional
information.
Question:
```

Topic sampling template.

```
{PASSAGE}
Please read the above Passage and summarize a
Topic it includes. Output the Topic directly
without any additional information.
Topic:
```

Topic-aware template (used across datasets).

```
{PASSAGE}
Please read the above Passage, and write a
Question related to "{TOPIC}" that a dense
retrieval model could use to find this Passage.
Directly output the Question without any
additional information.
Question:
```

D Per-Encoder Results for Naive Expansion

Table 8 reports the per-encoder breakdown of the Naive Expansion comparison summarized in Table 4. vMF-HSE outperforms its Naive counterpart on every encoder–dataset combination except SciFact, and consistently exhibits lower run-to-run variance.

E Generalization to Additional BEIR Domains

To assess zero-shot generalization beyond the five datasets in the main experiments, we evaluate on two CQADupStack sub-benchmarks of BEIR — Webmasters (technical web development Q&A, ~17K documents) and English (English language and usage Q&A, ~40K documents) — whose domains differ from the scientific, financial, and argumentative settings of the original five. All variants are evaluated under the same setting as Table 2 ($K=5$, $N=300$).

As shown in Tables 9 and 10, vMF achieves the highest average NDCG@10 on Webmasters (0.4288 vs. GMM_{raw} 0.4125 / $GMM_{\text{normalized}}$ 0.4158) and on CQADupStack-English (0.4982 vs. 0.4719 / 0.4909), with consistent improvements over the single-vector baseline. The pattern observed in the main experiments — that vMF mixture modeling captures multi-faceted document semantics more effectively than Euclidean alternatives — is therefore reproduced on domains and corpus scales beyond the original five datasets, supporting the framework’s zero-shot generalization.

Table 8: **Naive Expansion** ($N=300$ raw queries) vs. **vMF-HSE** ($K=5$) per encoder — **NDCG@10**. Mean $\pm$ standard deviation over 10 independent runs with different random seeds for LLM-based query generation.

Model	FiQA	ArguAna	SciDocs	SciFact	NFCorpus	Avg.
Naive _{Sentence-T5}	0.468 $\pm$ 0.015	0.383 $\pm$ 0.012	0.123 $\pm$ 0.018	0.525 $\pm$ 0.014	0.487 $\pm$ 0.016	0.397 $\pm$ 0.015
Naive _{BGE-M3}	0.486 $\pm$ 0.012	0.400 $\pm$ 0.010	0.132 $\pm$ 0.014	0.673 $\pm$ 0.012	0.557 $\pm$ 0.011	0.450 $\pm$ 0.012
Naive _{MiniLM}	0.473 $\pm$ 0.011	0.390 $\pm$ 0.009	0.150 $\pm$ 0.015	0.683 $\pm$ 0.011	0.553 $\pm$ 0.013	0.450 $\pm$ 0.012
Avg. Naive	0.476 $\pm$ 0.013	0.391 $\pm$ 0.010	0.135 $\pm$ 0.016	0.627 $\pm$ 0.012	0.532 $\pm$ 0.013	0.432 $\pm$ 0.013
vMF _{Sentence-T5}	0.492 $\pm$ 0.011	0.419 $\pm$ 0.009	0.165 $\pm$ 0.013	0.519 $\pm$ 0.010	0.517 $\pm$ 0.012	0.421 $\pm$ 0.011
vMF _{BGE-M3}	0.524 $\pm$ 0.008	0.424 $\pm$ 0.007	0.183 $\pm$ 0.006	0.651 $\pm$ 0.009	0.575 $\pm$ 0.008	0.471 $\pm$ 0.008
vMF _{MiniLM}	0.517 $\pm$ 0.006	0.421 $\pm$ 0.005	0.207 $\pm$ 0.008	0.672 $\pm$ 0.007	0.558 $\pm$ 0.009	0.473 $\pm$ 0.007
Avg. vMF (Ours)	0.511 $\pm$ 0.008	0.421 $\pm$ 0.007	0.185 $\pm$ 0.009	0.614 $\pm$ 0.009	0.550 $\pm$ 0.010	0.455 $\pm$ 0.009

Table 9: **CQADupStack-Webmasters** — **NDCG@10**.

Method	Sentence-T5	MiniLM	BGE-M3	Avg.
Baseline	0.3591	0.4158	0.3744	0.3831
GMM _{raw}	0.3876	0.4399	0.4101	0.4125
GMM _{normalized}	0.4030	0.4321	0.4122	0.4158
vMF (Ours)	0.4135	0.4437	0.4293	0.4288

Table 10: **CQADupStack-English** — **NDCG@10**.

Method	Sentence-T5	MiniLM	BGE-M3	Avg.
Baseline	0.3895	0.5025	0.4077	0.4332
GMM _{raw}	0.4312	0.5176	0.4669	0.4719
GMM _{normalized}	0.4378	0.5312	0.5038	0.4909
vMF (Ours)	0.4597	0.5413	0.4935	0.4982

F End-to-End Indexing Cost and Patience Sensitivity

F.1 End-to-End Per-Document Indexing Cost

The offline pipeline of vMF-HSE consists of three stages per document — (1) LLM-based query generation, (2) query encoding, and (3) vMF mixture fitting — each with distinct cost characteristics. Table 11 reports end-to-end measurements on NFCorpus (BGE-M3, single A100 GPU).

Table 11: **NFCorpus** — **per-document offline indexing cost (BGE-M3, A100)**. Encoding and vMF fitting times are measured; LLM generation cost scales linearly with the number of generated queries.

Method	Queries /doc	Tokens /doc	Enc. (s)	vMF fit (s)	Total (s)	Speedup	nDCG retention
Fixed- $N=300$	293.92	7,547	0.540	0.239	0.779	1.0 $\times$	100%
κ -Stop ($\tau=0.01$)	112.15	2,880	0.206	0.239	0.445	1.75 $\times$	99.89%
κ -Stop ($\tau=0.02$)	69.32	1,780	0.127	0.239	0.366	2.13 $\times$	100.18%
κ -Stop ($\tau=0.05$)	37.64	967	0.069	0.239	0.308	2.53 $\times$	98.30%
κ -Stop ($\tau=0.10$)	32.18	826	0.059	0.239	0.298	2.61 $\times$	98.21%

LLM query generation scales linearly with the number of generated queries: κ -Stop reduces the query count from 293.92 to 32.18 — an 89% reduction at $\tau=0.10$ — which translates directly into the same proportion of LLM generation cost saved.

Query encoding also scales linearly, dropping from 0.540s to 0.059s, while vMF fitting is essentially constant at ~ 0.24 s per document. End-to-end, κ -Stop reduces per-document offline cost by $1.75\times$ to $2.61\times$ while retaining 98.2–100.2% of the Fixed- $N=300$ retrieval performance. Since vMF-HSE is training-free, all three stages are performed offline at indexing time and incur no cost during online retrieval.

F.2 Sensitivity to the Patience Parameter P

While the convergence threshold τ sets the magnitude threshold for “no meaningful change,” the patience parameter P controls the temporal robustness of that judgment. Table 12 reports a sensitivity analysis on NFCorpus (MiniLM-L12-v2, $\tau=0.02$, $K=5$).

Table 12: **NFCorpus** — **effect of patience P on κ -Stop** (MiniLM-L12-v2, $\tau=0.02$). Retention is relative to the full Fixed- $N=294$ baseline (0.5565).

P	NDCG@10	Retention	Avg. Queries (out of 294)	
			Saving	
1	0.5466	98.22%	19	93.6%
2	0.5481	98.49%	28	90.4%
3 (default)	0.5459	98.10%	35	88.1%
4	0.5498	98.80%	49	83.4%
5	0.5486	98.58%	62	79.0%
6	0.5533	99.42%	77	73.8%
7	0.5600	100.63%	94	68.1%
8	0.5493	98.71%	110	62.5%

Retrieval quality is stable across the tested range, with NDCG@10 fluctuating within a narrow band of 0.5459–0.5600 (98.1–100.6% retention), and no value of P causes the algorithm to terminate prematurely or degrade meaningfully. The average number of queries grows monotonically with P — from 19 at $P=1$ to 110 at $P=8$ — reflecting the trade-off that larger patience permits more exploration before termination at higher computational cost, while the savings remain substantial across

the entire range (62.5–93.6%). Notably, larger P does not guarantee higher retrieval performance: NDCG@10 fluctuates non-monotonically in P . Across datasets with varying document lengths and corpus characteristics, $P=3$ consistently provided a reliable balance between performance retention and query saving, which is why we adopt it as the default; practitioners can safely use $P=3$ across diverse settings without per-dataset tuning.