RemoteRAG: A Privacy-Preserving LLM Cloud RAG Service

Yihang Cheng[1], Lan Zhang[1,2], Junyang Wang[1], Mu Yuan[3], Yunhao Yao[1]
[1]University of Science and Technology of China, Hefei, China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, China
[3]The Chinese University of Hong Kong, Hone Kong, China
yihangcheng@mail.ustc.edu.cn, zhanglan@ustc.edu.cn, iswangjy@mail.ustc.edu.cn
muyuan@cuhk.edu.hk, sa21011190@mail.ustc.edu.cn

Abstract

Retrieval-augmented generation (RAG) improves the service quality of large language models by retrieving relevant documents from credible literature and integrating them into the context of the user query. Recently, the rise of the cloud RAG service has made it possible for users to query relevant documents conveniently. However, directly sending queries to the cloud brings potential privacy leakage. In this paper, we are the first to formally define the privacy-preserving cloud RAG service to protect the user query and propose RemoteRAG as a solution regarding privacy, efficiency, and accuracy. For privacy, we introduce (n, ϵ)-DistanceDP to characterize privacy leakage of the user query and the leakage inferred from relevant documents. For efficiency, we limit the search range from the total documents to a small number of selected documents related to a perturbed embedding generated from (n, ϵ)-DistanceDP, so that computation and communication costs required for privacy protection significantly decrease. For accuracy, we ensure that the small range includes target documents related to the user query with detailed theoretical analysis. Experimental results also demonstrate that RemoteRAG can resist existing embedding inversion attack methods while achieving no loss in retrieval under various settings. Moreover, RemoteRAG is efficient, incurring only 0.67 seconds and 46.66KB of data transmission (2.72 hours and 1.43 GB with the non-optimized privacy-preserving scheme) when retrieving from a total of 10^5 documents.

1 Introduction

Large language models (LLMs) have attracted widespread attention since the release of ChatGPT (OpenAI, 2022a). However, LLM is not without its flaws. One major issue is its tendency to generate factually incorrect or purely fictional responses, a phenomenon known as hallucination (Leiser et al., 2024; Yao et al., 2023).

To mitigate this problem, retrieval-augmented generation (RAG) (Lewis et al., 2020) has been proposed to offer credible external knowledge, providing significant convenience for numerous tasks. RAG aims to understand the input query, extract relevant information from external data sources, and enhance the quality of the generated answers (Borgeaud et al., 2022; Lewis et al., 2020; Li et al., 2023b). Specifically, RAG allows the retrieval of relevant documents which can help understand or answer the input query and inserts them into the context of prompts to improve the output of LLM (Izacard and Grave, 2021). Its ability to enable LLM to provide answers with credible literature makes RAG an important technique in the application of LLM, leading to the development of many excellent and user-friendly open-source RAG projects (LangGenius, 2023; QuivrHQ, 2023; Chatchat-Space, 2023).

To leverage the power of RAG, a new concept RAG-as-a-Service (RaaS) has been proposed, gathering significant attention (Geniusee, 2024; Nuclia, 2024). In RaaS, the RAG service is entirely hosted online in the cloud. The user submits requests to the cloud with input queries to receive responses from RaaS. In this scenario, the cloud serves as the maintainer of the RAG service. While the current solution facilitates the wide adoption of RaaS, it raises serious privacy concerns. The input query may contain sensitive information, such as health conditions and financial status. Unfortunately, this data is not protected and must be uploaded in plaintext to the cloud in order to retrieve relevant documents. In this study, we aim to tackle a challenging question: *How can we minimize privacy leakage in queries for RaaS while ensuring the accuracy of the responses, all with minimal additional costs?*

Targeting the question above, we design a novel solution RemoteRAG. For privacy, we propose (n, ϵ)-DistanceDP inspired by differential privacy and an embedding perturbation mechanism, so that

Findings of the Association for Computational Linguistics: ACL 2025, pages 3820–3837
July 27 - August 1, 2025

the user can control the privacy leakage with a privacy budget ϵ in n-dimensional space of embeddings. We further study the potential privacy leakage in averaging the most relevant embeddings and find it within the constraint of (n, ϵ)-DistanceDP most of the time. For efficiency, we limit the search range from the total documents to a small number of selected documents, which are the relevant documents to a perturbed embedding generated from (n, ϵ)-DistanceDP. This small search range can save a significant amount of computation and communication costs used for privacy protection. For accuracy, we theoretically analyze the minimum size of the relevant documents to the perturbed embedding to ensure that they do contain target documents for the original query.

Contributions of this paper are listed as follows:

- To the best of our knowledge, we are the first to address the privacy-preserving cloud RAG service problem. We formally define the privacy-preserving cloud RAG service and characterize its corresponding threat model.
- We propose RemoteRAG as a solution to the privacy-preserving cloud RAG service regarding privacy, efficiency, and accuracy. We define (n, ϵ)-DistanceDP to characterize privacy leakage of the user query and design a mechanism to generate a perturbed embedding for the cloud for privacy, as well as to retrieve relevant documents within a minimum search range for efficiency. Accuracy is ensured by theoretical analysis of the minimum range produced by the perturbed embedding.
- We conduct extensive experiments to demonstrate that RemoteRAG can resist existing embedding inversion attack methods while achieving no loss in retrieval under various settings. The experiment results also show the efficiency of RemoteRAG, incurring only 0.67 seconds and 46.66KB of data transmission (2.72 hours and 1.43 GB with the non-optimized privacy-preserving scheme) when retrieving from a total of 10^6 documents.

2 Problem Formulation

We first formally define the problem in developing a privacy-preserving LLM cloud RAG service. The main notations used in this paper are listed in Table 1 for ease of reference.

Table 1: Notation table.

Sym.	Description
N	Number of RAG documents in the cloud
ϵ	Privacy budget
e_k	User query embedding
k	Number of top documents related to e_k
$e_{k'}$	Perturbed embedding
k'	Number of top documents related to $e_{k'}$
n	Dimensional space of embeddings

2.1 Problem Setup and Threat Model

One RAG request process involves two sides: a cloud and a user. The cloud hosts a substantial number (N) of documents as a RAG service. The user submits a request with a user query, and the RAG service in the cloud should retrieve top k relevant documents. An embedding model, shared between two sides, enables the user to convert the query into an embedding e_k. Additionally, the user has a privacy budget ϵ intended to measure and limit the privacy leakage of the query.

Threat model. We consider this scenario semi-honest, where both sides adhere to the protocol but the cloud is curious about the private information of the user query. During the RAG request process, the user should not reveal the semantic information of the query beyond the privacy budget ϵ allows. Apart from the query itself, we should also consider the protection of the query embedding and the indices of top k documents:

▹ *Protection of the query embedding.* Existing attack methods (Morris et al., 2023; Zhuang et al., 2024) have demonstrated that the semantic information can be extracted from the embedding if the embedding model is accessible. Consequently, safeguarding the semantic information of the query necessitates the protection of its embedding as well.

▹ *Protection of the indices of top k documents.* The embeddings of top k documents are situated in proximity to the query embedding, which potentially leads to the leakage of the query embedding. Specifically, the average of top k document embeddings could be close to the query embedding, for which privacy leakage should be carefully studied.

2.2 Design Scope

In RemoteRAG, we examine the potential privacy leakage occurring during the data transmission between the cloud and the user. Considerations of offline RAG services downloaded directly from the cloud or internal privacy issues (Zeng et al., 2024) specific to RAG are beyond the scope of this paper.

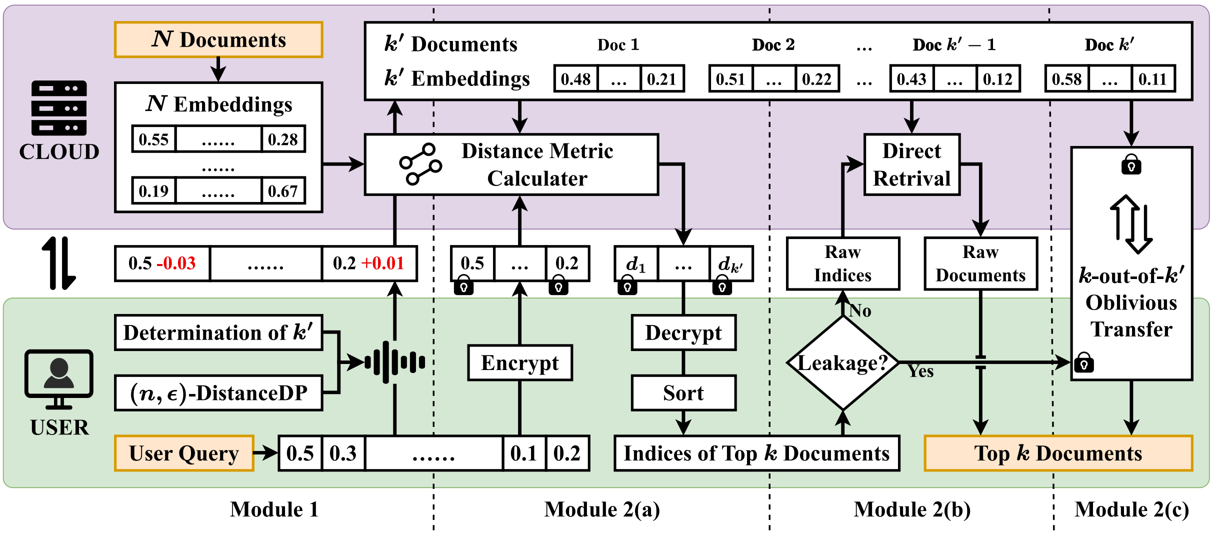


Figure 1: The flowchart of RemoteRAG. Module 1 preserves privacy with (n, ϵ)-DistanceDP and improves efficiency by limiting the search range. Module 2 retrieves documents with different choices based on leakage circumstances.

3 System Design

3.1 Overview

The flowchart of RemoteRAG is shown in Figure 1. Under the control of (n, ϵ)-DistanceDP, module 1 aims to reduce the search range to enhance efficiency while ensuring accuracy. Module 2 targets safely retrieving top k relevant documents within the limited range from two optional choices under different leakage circumstances.

3.2 Range Limitation with Privacy Budget

3.2.1 Generation of Perturbation

Given the requirement that the query embedding should not be transmitted to the cloud, we opt to send a perturbed embedding instead. To assess the potential privacy leakage in the perturbed embedding, we utilize the differential privacy (DP) theory to define (n, ϵ)-DistanceDP in n-dimensional space:

Definition 1 ((n, ϵ)-DistanceDP)**.** A mechanism K satisfies (n, ϵ)-DistanceDP if and only if $\forall x, x' \in \mathbb{R}^n$:

$$L(K(x), K(x')) \le \epsilon \|x - x'\|$$

where ϵ is a given privacy budget, $\|x - x'\|$ denotes L2 distance, and $L(K(x), K(x')) = \ln \frac{Pr(K(x)=y)}{Pr(K(x')=y)}$ represents the distance between the probabilities of any target value y drawn from the distributions $K(x)$ and $K(x')$ generated by points x and x'.

To apply (n, ϵ)-DistanceDP in RemoteRAG, for any query embedding e_k, the user generates a perturbed embedding $e_{k'}$ using a noise function, which should satisfy that from the perspective of $e_{k'}$, the probability of generating the query embedding e_k and another random embedding e_x around $e_{k'}$ with the noise function differs by at most a multiplicative factor of $e^{-\epsilon \|e_k - e_x\|}$.

Noise function. The property above can be achieved by utilizing the Laplace distribution, as discussed in (Andrés et al., 2013; Dwork et al., 2006). The primary difference lies in our application within higher n-dimensional space. Given the privacy budget $\epsilon \in \mathbb{R}^+$ and the actual point $x_0 \in \mathbb{R}^n$, the probability density function (pdf) of the noise function at any other point $x \in \mathbb{R}^n$ is given by:

$$D_{n,\epsilon}(x|x_0) \propto e^{-\epsilon \|x - x_0\|}$$

Pratical generation guideline. Directly generating a point according to the above distribution is challenging. Therefore, we separately generate the radial component and direction vector:
- Radial component $r = \|x - x_0\|$. Its marginal distribution is given by $D_{n,\epsilon}(r) \propto r^{n-1} e^{-\epsilon r}$, which corresponds exactly to the pdf of the gamma distribution with shape parameter n and scale parameter $\frac{1}{\epsilon}$. Therefore, $r \sim D_{n,\epsilon}(r) = \text{Gamma}(n, \frac{1}{\epsilon})$.
- Direction vector $\mathbf{v} = \{v_1, \cdots, v_n\}$. It should be sampled from a uniform distribution on the n-dimensional unit sphere. This can be accomplished by independently sampling $t_i \sim N(0, 1)$ from the standard normal distribution and then normalizing it (Marsaglia, 1972): $v_i = \frac{t_i}{\sqrt{\sum_{j=1}^n t_j^2}}, i \in [1, n]$.

From r to ϵ: $\epsilon \approx \frac{n}{r}$. Apart from drawing r from the gamma distribution formed by ϵ, we can also estimate the value of the privacy budget ϵ from a

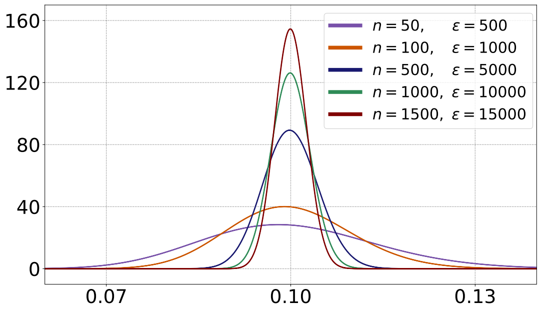
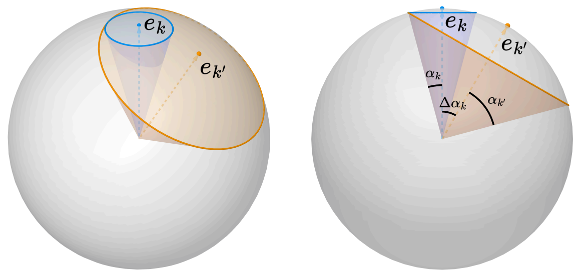


Figure 2: The probability density function of different gamma distributions within $[0.06, 0.14]$ range.



(a) Oblique projection. Top k' documents related to $e_{k'}$ include top k documents related to e_k.

(b) Orthographic projection. $\alpha_{k'} = \alpha_k + \Delta\alpha_k$.

Figure 3: Illustration in 3-dimensional projection.

given perturbation r. Since $r \sim \text{Gamma}(n, \frac{1}{\epsilon})$, the expected value of it is $\bar{r} = \frac{n}{\epsilon}$. We notice that current embedding models often produce embeddings with a large dimension (e.g., 384 for all-MiniLM-L12-v2 (SentenceTransformers, 2021), 768 for gtr-t5-base (SentenceTransformers, 2022), and 1536 for text-embedding-ada-002 (OpenAI, 2022b)). The pdf of $\text{Gamma}(n, \frac{1}{\epsilon})$ becomes increasingly steep as the dimension n increases, as illustrated in Figure 2. This implies that the radial components drawn from the distribution are likely to cluster around $\bar{r}$. Therefore, for any given perturbation r, the privacy budget $\epsilon \approx \frac{n}{r}$.

3.2.2 Calculation of Search Range

After generating the perturbation, the user then requests the cloud to retrieve top k' documents related to the perturbed embedding $e_{k'}$, thereby limiting the search range from N to k'. To maintain accuracy, it is crucial to ensure that these k' documents include top k documents related to the query embedding e_k. This requirement motivates the need to determine the appropriate value for k'.

Lemma 1. Assume that there are N embeddings uniformly distributed on the surface of the n-dimensional unit sphere. Let α_k be the polar angle of the surface area formed by top k embeddings related to any given embedding. Then, k and α_k satisfy the following relationship:

$$k = N \cdot \frac{\Omega_{n-1}(\pi)}{\Omega_n(\pi)} \cdot \int_0^{\alpha_k} \sin^{n-2}\theta \, d\theta$$

where $\Omega_n(\pi) = \frac{2\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2})}$ represents the surface area of the unit n-sphere.

From Lemma 1, we can derive the polar angle α_k from k. Since the perturbation is small ($r \ll 1$), the perturbed angle $\Delta\alpha_k$ between e_k and $e_{k'}$ can be approximated as r. To ensure that top k' documents related to $e_{k'}$ include top k documents related to e_k, we further propose Theorem 1 and illustrate the principle in Figure 3.

Theorem 1. Under the conditions specified in Lemma 1, given two embeddings e_k and $e_{k'}$ with the perturbed angle $\Delta\alpha_k$, to ensure that top k' embeddings related to $e_{k'}$ include top k embeddings related to e_k, k' and k satisfy the following relationship:

$$\Delta k = k' - k = N \cdot \frac{\Omega_{n-1}(\pi)}{\Omega_n(\pi)} \cdot \int_{\alpha_k}^{\alpha_{k'}} \sin^{n-2}\theta \, d\theta$$

where $\alpha_{k'} = \alpha_k + \Delta\alpha_k$.

Choosing an appropriate value for ϵ. In practice, the user usually has clear privacy and cost thresholds, which give the upper and lower bounds of the privacy budget ϵ, respectively:

- The privacy threshold gives the upper bound. As the perturbation r increases, the attack performance will gradually decrease (Figure 4(a)) below the threshold set by the user. Since $\epsilon \approx \frac{n}{r}$, the minimum value of r determines the maximum value of ϵ.
- The cost threshold gives the lower bound. The computation and communication costs increase as k' increases, as shown in Figures 5(a) and 5(b). The maximum value of k' determines the maximum value of the perturbed angle $\Delta\alpha_k$ (Theorem 1), which is approximated as r and determines the minimum value of ϵ.

3.3 Retrieval with Cryptographic Protection

3.3.1 Homomorphic Encryption for Indices

After limiting the search range to k' documents, we now focus on obtaining the indices of top k documents for the query embedding e_k. Based on documentations of several open-source vector databases such as ChromaDB (Chroma, 2022), FAISS (Douze et al., 2024), and Elasticsearch (Elastic, 2010), we find that only two distance metrics are being used by all of them and set to be default: L2 distance and cosine distance.

Definition 2 (Distance metrics)**.** Given two normalized embeddings e_a and e_b of the same dimension, L2 distance and cosine distance are calculated as follows:

$$d_{l2}(e_a, e_b) = \|e_a - e_b\|$$

$$d_{\cos}(e_a, e_b) = 1 - \frac{\langle e_a, e_b \rangle}{\|e_a\| \cdot \|e_b\|} = 1 - \langle e_a, e_b \rangle$$

Further analysis in Theorem 2 reveals that using which distance metric does not affect the ranking of normalized vectors. Therefore, we only consider cosine distance as the standard metric in this paper.

Theorem 2. Given two normalized embeddings e_a and e_b of the same dimension, L2 distance and cosine distance have the following relationship:

$$d_{l2}(e_a, e_b) = \sqrt{2 d_{\cos}(e_a, e_b)}$$

Recall that the query embedding e_k should not be revealed to the cloud. In this typical secure multi-party computation scenario, considering that secret sharing usually needs at least three non-colluding parties or to introduce a trustful third party, which is a strong assumption and may not be practical, we leverage homomorphic encryption in RemoteRAG. Because cosine distance involves only linear operations, we propose using partially homomorphic encryption. Compared to fully homomorphic encryption, it is more computationally efficient and sufficient for calculating cosine distance.

Specifically, the user encrypts the query embedding e_k and sends the encrypted form $[\![e_k]\!]$ to the cloud. The cloud then calculates cosine distances $[\![d_{\cos}(e_k, e)]\!] = d_{\cos}([\![e_k]\!], e)$ in encrypted form between $[\![e_k]\!]$ and each document embedding e related to $e_{k'}$. Upon receiving these distances, the user decrypts them and sorts the results to find the indices of top k documents related to the query.

3.3.2 Document Retrieval with Indices

We then carefully analyze whether these indices are safe to be directly sent to the cloud to retrieve the corresponding documents.

If the cloud receives these indices, it also means that it knows which k documents are the closest to the user query. The cloud can average these document embeddings to construct an embedding $\bar{e}$ which approximates the query embedding e_k. Therefore, we should measure how close the two embeddings e_k and $\bar{e}$ are.

Theorem 3. Given a target query embedding e_k and the mean embedding $\bar{e}$ of top k relevant document embeddings, the mean angle ω between e_k and $\bar{e}$ satisfies

$$\tan \omega = \frac{\tan \alpha_k}{\sqrt{k}}$$

where α_k is calculated from Lemma 1.

From Theorem 3, we characterize the approximation between e_k and $\bar{e}$ with the mean angle ω. Recall that the privacy budget ϵ generates a perturbation with a mean of $\frac{n}{\epsilon}$ (i.e. $\Delta\alpha_k \approx \bar{r} = \frac{n}{\epsilon}$), as described in Section 3.2.1. Based on the leakage circumstances, we offer two choices to retrieve target documents:

- Direct retrieval from indices. If $\omega \geq \Delta\alpha_k$, $\bar{e}$ is within the control of the privacy budget ϵ and the indices require no extra protection. The user can directly send the indices to the cloud to retrieve the corresponding documents.
- Safe retrieval with k-out-of-k' oblivious transfer (OT). If $\omega < \Delta\alpha_k$, $\bar{e}$ is even closer to the query embedding e_k than the perturbed embedding $e_{k'}$, therefore requiring further protection. In this situation, we suggest using the k-out-of-k' OT protocol (Chou and Orlandi, 2015), which allows a sender (the cloud) with a set of k' messages (documents) to transfer a subset of k messages to a receiver (the user) while remaining oblivious to the specific subset (indices) chosen by the receiver. The OT protocol is described in Appendix A.1.

4 Analysis on RemoteRAG

4.1 Security Analysis

RemoteRAG must adhere to the privacy-preserving goal outlined in the threat model in Section 2.1. The cloud receives the following messages which may leak the semantic meaning of the user query:

Module 1. The cloud receives the perturbed embedding. The perturbation r between the user embedding and the perturbed embedding is sampled from $\text{Gamma}(n, \frac{1}{\epsilon})$, which satisfies (n, ϵ)-DistanceDP, controlled by the privacy budget ϵ.

Module 2(a). The cloud receives the encrypted form of the query embedding. Without the secret key, the cloud cannot reverse engineer the query embedding. The computation of cosine distances is guaranteed by PHE, which does not leak any information either.

Table 2: Comparison among RemoteRAG, privacy-ignorant and privacy-conscious services.

		Security	Communication		
		User Query	Rounds	Numbers (β units)	Documents (η units)
Privacy-ignorant Service		✗	1	n	k
Privacy-conscious Service		✓	2	$n + 2N + 1$	N
RemoteRAG	(Direct)	(n, ϵ)-DistanceDP	2	$2n + k + k' + 1$	k
	(OT)			$2(n + k' + 1)$	k'

Module 2(b), if $\omega \geq \Delta\alpha_k$. The cloud receives the indices of top k relevant documents. The perturbation from the mean embedding of top k relevant document embeddings is within the protection scope of the privacy budget ϵ.

Module 2(c), if $\omega < \Delta\alpha_k$. The cloud and the user perform the k-out-of-k' OT protocol. The property of the OT protocol ensures that the indices of top k relevant documents are not visible to the cloud.

From the analysis above, we demonstrate that the user receives top k documents without disclosing any information about the user query under the constraint of the given privacy budget ϵ.

4.2 Communication Analysis

We analyze communication from two aspects: the number of communication rounds and the size of communication. We define one communication round as the transmission of a message from one side to another and back to the original side. The size of one number and one document are set to be β units and η units, respectively.

Module 1. There is one message transmitted from the user to the cloud (0.5 communication round), containing the perturbed embedding $e_{k'}$ and the corresponding k'. $e_{k'}$ is a vector of length n with $n\beta$ units, while k' is a number occupying β units.

Module 2(a). There is 1 communication round. First, the encrypted form $[\![e_k]\!]$ of the query embedding is sent to the cloud, occupying $n\beta$ units. Second, the cloud sends back encrypted cosine distances, occupying $k'\beta$ units.

Module 2(b), if $\omega \geq \Delta\alpha_k$. There is 1 communication round. The user sends the indices ($k\beta$ units) to the cloud and the cloud returns the target documents ($k\eta$ units).

Module 2(c), if $\omega < \Delta\alpha_k$. There are 1.5 communication rounds and $(k'+1)\beta + k'\eta$ units of messages for the k-out-of-k' OT protocol. Details can be found in Appendix A.1.

By summing these, if $\omega \geq \Delta\alpha_k$, the total number of communication rounds is 2.5, and the size of communication is $(2n + k + k' + 1)\beta + k\eta$ units; if $\omega < \Delta\alpha_k$, the total number of communication rounds is 3, and the size of communication is $2(n + k' + 1)\beta + k'\eta$ units.

Practical optimization. The number of communication rounds can be further reduced in practice. For example, the user can simultaneously send both the perturbed embedding and the encrypted form of the query embedding to the cloud in a single communication to reduce 0.5 round for modules 1 and 2(a). Additionally, the cloud can send the encrypted cosine distances and start the OT protocol together to reduce another 0.5 round for modules 2(a) and 2(c). Therefore, no matter whether with module 2(b) or 2(c), the total number of communication rounds can be further reduced to 2.

4.3 Special Cases

The privacy-ignorant cloud RAG service. A privacy-ignorant cloud RAG service does not account for user query privacy, requiring the user to upload the query embedding and receive top k documents directly. This represents a special case of RemoteRAG, achieved by setting $\epsilon \rightarrow \infty$, with the perturbation $r \sim \text{Gamma}(n, 0)$ (i.e. no perturbation). The service requires 1 communication round with $n\beta + k\eta$ units in this case.

The privacy-conscious cloud RAG service. A privacy-conscious cloud RAG service aims to fully protect user query privacy. This can be regarded as the combination of modules 2(a) and 2(c) in RemoteRAG, where $k' = N$. This is another special case of RemoteRAG, achieved by setting $\epsilon \rightarrow 0$, with the perturbation $r \sim \text{Gamma}(n, \infty)$ and $k' = N$ (i.e. cryptographic computation over all N documents). This case requires 2 communication rounds with $(n + 2N + 1)\beta + N\eta$ units.

The comparison between RemoteRAG and these two special cases is presented in Table 2.

Table 3: Parameter settings for experiments. "\" means "Not Applicable" and "✓" means "Variable".

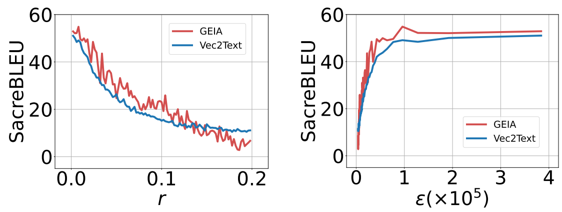
		Dataset (Size)	Embedding Model	k	$\epsilon/r/k'$
Privacy Study	Section 5.2	\	T5	\	\
	Appendix B.1	\	T5	\	\
Accuracy Study	Section 5.3	MS ($10^4/10^5/10^6$)	✓	✓	✓
	Appendix B.2	NQ (all), TQA (all), MS (all)	✓	✓	✓
Efficiency Study	Section 5.4	MS (10^5)	T5	5	✓
	Appendix B.3	MS ($10^5/10^7$)	T5	5	$k' = 160$

Table 4: Embedding dimensions of embedding models.

Embedding Model	Dimension
all-MiniLM-L12-v2 (MiniLM)	384
all-mpnet-base-v2 (MPNet)	768
gtr-t5-base (T5)	768
text-embedding-ada-002 (OpenAI-1)	1536
text-embedding-3-large (OpenAI-2)	3072

Table 5: Number of sentences in datasets.

Dataset	Sentences
Nature Questions (NQ)	26299
TriviaqQA (TQA)	847579
MS MARCO (MS)	1112939



(a) SacreBLEU-r (b) SacreBLEU-ϵ

Figure 4: The SacreBLEU metric corresponding to the perturbation r and the privacy budget ϵ.

5 Experiments

We evaluate RemoteRAG under various settings detailed in Section 5.1. Our key findings are:

▹ For privacy, RemoteRAG controls the semantic information leakage of the user query with the privacy budget. **[Section 5.2]**

▹ For accuracy, RemoteRAG achieves lossless document retrieval. **[Section 5.3]**

▹ For efficiency, RemoteRAG introduces little extra computation and communication costs while preserving privacy. **[Section 5.4]**

5.1 Experiment Setup

Embedding Model Details. We use three open-sourced embedding models: all-MiniLM-L12-v2 (MiniLM) (SentenceTransformers, 2021), all-mpnet-base-v2 (MPNet), gtr-t5-base (T5) (SentenceTransformers, 2022); and two OpenAI proprietary embedding models: text-embedding-ada-002 (OpenAI-1) (OpenAI, 2022b), text-embedding-3-large (OpenAI-2) (OpenAI, 2024) . The details of these embedding models are shown in Table 4.

Dataset Details. We use Natural Questions (NQ) (Kwiatkowski et al., 2019), TriviaqQA (TQA) (Joshi et al., 2017), and MS MARCO (MS) (Nguyen et al., 2016) as the RAG datasets. Sizes of the datasets are shown in Table 5.

Parameter Settings. We list the parameter settings of all experiments in Table 3. We group $\epsilon/r/k'$ in one column since they can deduce from each other. Additionally, the experimental results shown in all figures and tables are the averages of 50 independent experiments.

Environment. All of our experiments are conducted using PyTorch (Paszke et al., 2019) on an Ubuntu 22.04 server with two 28-core Intel(R) Xeon(R) Gold 5420+ processors and two Nvidia A40 48GB GPUs.

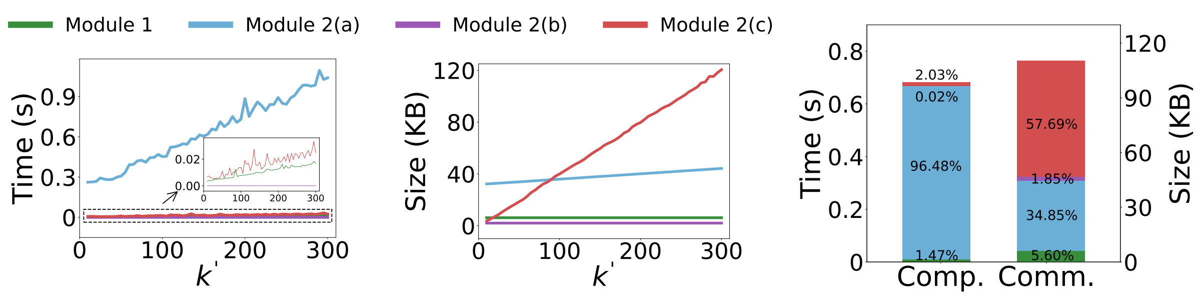
5.2 Privacy Study

We first examine privacy leakage and control with the privacy budget in RemoteRAG. We apply attack methods GEIA (Li et al., 2023a) and Vec2Text (Morris et al., 2023) and use SacreBLEU (Post, 2018) to measure the difference between the original query and the recovered query.

From an intuitive perspective, we plot the SacreBLEU metric against the perturbation r to see how the perturbation affects the attack. From the results in Figure 4(a), we observe that the attack performance drops from 50 to 10 as the perturbation

Table 6: RemoteRAG achieves no loss in retrieval under various settings in our experiments.

	N	k	r	Embedding Model
	$10^4/10^5/10^6$	5/10/15/20	0.03/0.05/0.07/0.1	MiniLM/MPNet/T5/OpenAI-1/OpenAI-2
Recall	100%	100%	100%	100%



(a) Computation cost.

(b) Communication cost.

(c) Cost breakdown ($k' = 160$).

Figure 5: Efficiency study of each module.

increases from 0 to 0.2. When the perturbation reaches 0.2, which is relatively large, the attack becomes completely ineffective. It demonstrates the effectiveness of adding a perturbation to the original query embedding for protection.

From another perspective, we analyze the variation in the attack performance against the privacy budget ϵ. The results are shown in Figure 4(b). Overall, the performance of the attack improves as ϵ increases. This is within our expectation, since a larger privacy budget means a looser tolerance for privacy leakage, which allows for a smaller perturbation and ultimately leads to a better attack performance in Figure 4(a). By setting an upper bound for the privacy budget, the user can control the privacy leakage of the perturbed embedding.

5.3 Accuracy Study

To demonstrate the correctness of theoretical analysis for the calculation of k', we conduct experiments under various settings: different total numbers N of documents, different numbers k of top relevant documents, different sizes r of the perturbation chosen by the user, and different embedding models. We use recall to evaluate the proportion of top k documents included in the results.

Throughout our experiment, we have not encountered a situation where any of the top k documents are missing from the set of k' documents. As shown in Table 6, recall in all settings is 100%, indicating that all top k documents are included in the set of k' documents computed in module 1 and therefore can be correctly selected by module 2.

Table 7: Efficiency comparison ($k' = 160$).

		Comp.	Comm.
Privacy-ignorant Service		`3.15ms`	`8.00KB`
Privacy-conscious Service		`2.72hr`	`1.43GB`
RemoteRAG	(Direct)	`0.67s`	`46.66KB`
	(OT)	`0.68s`	`108.24KB`

5.4 Efficiency Study

For efficiency, the metrics are running time for computation cost and transmission size for communication cost. We provide the results of each module in Figure 5. The linear results are consistent with the analysis in Table 2. We highlight $k' = 160, r = 0.03$ in Figure 5(c) with the attack performance moderate at around 30 shown in Figure 4, and compare the results with two baselines (see Section 4.3) in Table 7.

Computation cost. From the results in Figure 5(a), the most computationally intensive task occurs in module 2(a), which accounts for over 95% of the total computation cost. This substantial cost renders it impractical for scenarios involving a large number of documents when calculating cosine distances. As indicated in Table 7, the privacy-conscious service requires 2.72 hours in total to process a single user request, which is considered unacceptable. But in RemoteRAG, we only take less than 1 second for calculating cosine distances, due to the search range limitation in module 1, which saves huge computation cost.

Communication cost. From the results in Figure 5(b), module 2(a) has a larger starting point, but the transmission size of module 2(c) soon surpasses module 2(a) as k' increases. Basic parameters in PHE cause the former while the latter is due to the larger size of encrypted documents. When k' is relatively large, the transmission size becomes unacceptable. As shown in Table 7, the privacy-conscious service incurs a considerable transmission size (1.43GB) of N documents. Again, with the search range limitation in module 1, RemoteRAG only needs to transfer about 100 KB of data.

Direct and OT. We compare the results using module 2(b) or module 2(c) in Table 7. There is little increase in computation cost but a large increase in communication cost. The cost breakdown illustrated in Figure 5(c) provides an intuitive distribution of costs. OT does not bring much computation cost (from 0.02% to 2.03% compared to Direct), but its necessity to transfer k' encrypted documents increases the transmission size from 1.85% to 57.69%.

6 Related Work

RAG techniques. Many researches have improved the performance of RAG by exploring the potential in embedding model architectures (Li and Li, 2023; Voyage, 2024; Chen et al., 2024; OpenAI, 2022b, 2024), chunking strategies (LlamaIndex, 2023; LangChain, 2024; Sophia Yang, 2023), and query optimization (Zhou et al., 2023; Dhuliawala et al., 2023; Ma et al., 2023). They do not overlap with RemoteRAG and can be directly applied to improve the quality of RAG results.

RAG protection. Recently, Grislain (2024); Koga et al. (2024) focus on the differential protection solution to the leakage of private information in the retrieved documents to LLM. However, they are not protecting the user query, which is different from our goal in this paper.

Prompt protection. Privacy-preserving prompt engineering is a technique for LLM inference. Gupta et al. (2024); Kan et al. (2023); Chen et al. (2023) modify the prompt directly to remove sensitive information but fail to provide rigorous privacy protection. Their application in RemoteRAG needs careful investigation in the change of the query embedding to avoid accuracy decrease. Tang et al. (2024); Hong et al. (2024) aim to protect the context of the prompt, therefore cannot be applied in RemoteRAG.

7 Conclusion

In this paper, we are the first to address and formally define the privacy-preserving cloud RAG service problem. We propose RemoteRAG as a solution regarding privacy, efficiency, and accuracy. (n, ϵ)-DistanceDP is introduced to characterize the privacy leakage of the user query. The perturbation limits the search range, significantly saving computation and communication costs. Theoretical analysis ensures the accuracy. Experimental results also demonstrate the superiority of RemoteRAG in privacy, efficiency, and accuracy, compared to privacy-ignorant and privacy-conscious services.

Acknowledgment

Lan Zhang is the corresponding author. This research was supported by the National Key R&D Program of China 2021YFB2900103, China National Natural Science Foundation with No. 62441228 and No. 623B2093, Science and Technology Tackling Program of Anhui Province, No. 202423k09020016.

Ethical Considerations

No disclosure risk. The privacy-preserving cloud RAG service is a new scenario proposed in this paper, which has not been formally used in practice. RemoteRAG as the first solution to the potential privacy issues in this scenario can promote the development of this field with no disclosure risk.

Open-sourced content in experiments. The open-sourced models and datasets used in our experiments are all downloaded from HuggingFace without modification. We believe that using them appropriately according to their original purpose will not have a direct negative impact.

Compliance with laws and regulations. RemoteRAG is proposed as a solution to potential privacy leakage in the privacy-preserving cloud RAG service, making it compliant with laws and regulations such as GDPR (Voigt and von dem Bussche, 2024).

Limitations

Limitation of PHE. PHE supports only the addition operation. This restricts the variety of similarity distances RemoteRAG can calculate. For example, FAISS offers to use Lp and Jaccard metrics, which may not be easy to use PHE. Besides, RAG may also be combined with keyword searching for better retrieval results. These require further investigation.

Proprietary embedding model. Although open-source embedding models have already achieved great performance, the cloud may still consider using its own proprietary embedding model. In this scenario, the user cannot calculate the query embedding locally and therefore cannot directly generate the perturbation for protection.

▹ Some studies (Hao et al., 2022; Hou et al., 2023) have explored using fully homomorphic encryption on Transformer architecture models. But they still suffer from huge computation costs.

▹ Another possible solution might be to perturb the query itself. From theoretical analysis of how the perturbation to the original query affects the output embeddings of a proprietary embedding model, we can establish the relationship between the size of query perturbation and the size of embedding perturbation. In this way, it effectively completes the generation of perturbation in Section 3.2.1 from another perspective without introducing any additional cost, and seamlessly connects to the design of RemoteRAG. However, the theoretical analysis still remains a challenging problem.

References

Miguel E. Andrés, Nicolás Emilio Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. 2013. Geo-indistinguishability: differential privacy for location-based systems. In *2013 ACM SIGSAC Conference on Computer and Communications Security, CCS'13, Berlin, Germany, November 4-8, 2013*, pages 901–914. ACM.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Chatchat-Space. 2023. Langchain-Chatchat. `https://github.com/chatchat-space/Langchain-Chatchat`.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *CoRR*, abs/2402.03216.

Yu Chen, Tingxin Li, Huiming Liu, and Yang Yu. 2023. Hide and seek (has): A lightweight framework for prompt privacy protection. *CoRR*, abs/2309.03057.

Tung Chou and Claudio Orlandi. 2015. The simplest protocol for oblivious transfer. In *Progress in Cryptology - LATINCRYPT 2015 - 4th International Conference on Cryptology and Information Security in Latin America, Guadalajara, Mexico, August 23-26, 2015, Proceedings*, volume 9230 of *Lecture Notes in Computer Science*, pages 40–58. Springer.

Chroma. 2022. Chroma. `https://github.com/chroma-core/chroma`.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *CoRR*, abs/2309.11495.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *CoRR*, abs/2401.08281.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer.

Elastic. 2010. Elasticsearch. `https://github.com/elastic/elasticsearch`.

Geniusee. 2024. RAG as a service. `https://geniusee.com/retrieval-augmented-generation`.

Nicolas Grislain. 2024. RAG with differential privacy. *CoRR*, abs/2412.19291.

Brij B. Gupta, Akshat Gaurav, Varsha Arya, Wadee Alhalabi, Dheyaaldin Alsalman, and Pandi Vijayakumar. 2024. Enhancing user prompt confidentiality in large language models through advanced differential encryption. *Comput. Electr. Eng.*, 116:109215.

Meng Hao, Hongwei Li, Hanxiao Chen, Pengzhi Xing, Guowen Xu, and Tianwei Zhang. 2022. Iron: Private inference on transformers. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Junyuan Hong, Jiachen T. Wang, Chenhui Zhang, Zhangheng Li, Bo Li, and Zhangyang Wang. 2024. DP-OPT: make large language model your privacy-preserving prompt engineer. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Xiaoyang Hou, Jian Liu, Jingyu Li, Yuhan Li, Wen-jie Lu, Cheng Hong, and Kui Ren. 2023. Ciphergpt: Secure two-party GPT inference. *IACR Cryptol. ePrint Arch.*, page 1147.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 874–880. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics.

Zhigang Kan, Linbo Qiao, Hao Yu, Liwen Peng, Yifu Gao, and Dongsheng Li. 2023. Protecting user privacy in remote conversational systems: A privacy-preserving framework based on text sanitization. *CoRR*, abs/2306.08223.

Tatsuki Koga, Ruihan Wu, and Kamalika Chaudhuri. 2024. Privacy-preserving retrieval augmented generation with differential privacy. *CoRR*, abs/2412.04697.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

LangChain. 2024. How to recursively split text by characters. `https://python.langchain.com/v0.2/docs/how_to/recursive_text_splitter/`.

LangGenius. 2023. Dify. `https://github.com/langgenius/dify`.

Florian Leiser, Sven Eckhardt, Valentin Leuthe, Merlin Knaeble, Alexander Mädche, Gerhard Schwabe, and Ali Sunyaev. 2024. HILL: A hallucination identifier for large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*, pages 482:1–482:13. ACM.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Haoran Li, Mingshi Xu, and Yangqiu Song. 2023a. Sentence embedding leaks more information than you expect: Generative embedding inversion attack to recover the whole sentence. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14022–14040. Association for Computational Linguistics.

Jiatong Li, Yunqing Liu, Wenqi Fan, Xiao-Yong Wei, Hui Liu, Jiliang Tang, and Qing Li. 2023b. Empowering molecule discovery for molecule-caption translation with large language models: A chatgpt perspective. *CoRR*, abs/2306.06615.

Xianming Li and Jing Li. 2023. Angle-optimized text embeddings. *CoRR*, abs/2309.12871.

LlamaIndex. 2023. Evaluating the Ideal Chunk Size for a RAG System using LlamaIndex. `https://www.llamaindex.ai/blog/evaluating-the-ideal-chunk-size-for-a-rag-system-using-llamaindex-6207e5d3fec5`.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting for retrieval-augmented large language models. *CoRR*, abs/2305.14283.

George Marsaglia. 1972. Choosing a point from the surface of a sphere. *The Annals of Mathematical Statistics*, 43(2):645–646.

John X. Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander M. Rush. 2023. Text embeddings reveal (almost) as much as text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12448–12460. Association for Computational Linguistics.

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. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Nuclia. 2024. Nuclia, the all-in-one RAG as a service platform. `https://nuclia.com/rag-as-a-service/`.

OpenAI. 2022a. Introducing ChatGPT. `https://openai.com/index/chatgpt/`.

OpenAI. 2022b. New and improved embedding model. `https://openai.com/index/new-and-improved-embedding-model/`.

OpenAI. 2024. New embedding models and API updates. `https://openai.com/index/new-embedding-models-and-api-updates/`.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191. Association for Computational Linguistics.

QuivrHQ. 2023. Quivr. https://github.com/QuivrHQ/quivr.

SentenceTransformers. 2021. all-MiniLM-L12-v2. https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2.

SentenceTransformers. 2022. gtr-t5-base. https://huggingface.co/sentence-transformers/gtr-t5-base.

Ph.D. Sophia Yang. 2023. Advanced RAG 01: Small-to-Big Retrieval. https://towardsdatascience.com/advanced-rag-01-small-to-big-retrieval-172181b396d4.

Xinyu Tang, Richard Shin, Huseyin A. Inan, Andre Manoel, Fatemehsadat Mireshghallah, Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, and Robert Sim. 2024. Privacy-preserving in-context learning with differentially private few-shot generation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Paul Voigt and Axel von dem Bussche. 2024. *The EU General Data Protection Regulation (GDPR)*. Springer.

Voyage. 2024. Embeddings. https://docs.voyageai.com/docs/embeddings.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Munan Ning, and Li Yuan. 2023. LLM lies: Hallucinations are not bugs, but features as adversarial examples. *CoRR*, abs/2310.01469.

Shenglai Zeng, Jiankun Zhang, Pengfei He, Yue Xing, Yiding Liu, Han Xu, Jie Ren, Shuaiqiang Wang, Dawei Yin, Yi Chang, and Jiliang Tang. 2024. The good and the bad: Exploring privacy issues in retrieval-augmented generation (RAG). *CoRR*, abs/2402.16893.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Shengyao Zhuang, Bevan Koopman, Xiaoran Chu, and Guido Zuccon. 2024. Understanding and mitigating the threat of vec2text to dense retrieval systems. *CoRR*, abs/2402.12784.

Algorithm 1: RemoteRAG: Module 1

Module 1:
    // user side
    generate a random perturbation $r \sim \text{Gamma}(n, \frac{1}{\epsilon})$
    generate its direction vector $\mathbf{v} = \{v_1, \cdots, v_n\}$, where $v_i = \frac{t_i}{\sqrt{\sum_{j=1}^{n} t_j^2}}, t_i \in \mathcal{N}(0, 1), i \in [1, n]$
    compute the perturbed embedding $e_{k'} = e_k + r\mathbf{v}$
    determine k' from Theorem 1
    // cloud side
    retrieve top k' documents related to $e_{k'}$

Algorithm 2: RemoteRAG: Module 2

Module 2:
    // user side
    encrypt $[\![e_k]\!]$ using PHE
    // cloud side
    foreach $e \in$ document embeddings retrieved from Algorithm 1 **do**
        calculate cosine distance in encrypted form $[\![d_i]\!] = d_{\cos}([\![e_k]\!], e)$
    // user side
    decrypt cosine distances $d_i, i \in [1, k']$
    sort cosine distances to obtain the indices of top k documents related to e_k
    if $\arctan \frac{\tan \alpha_k}{\sqrt{k}} \geq \frac{n}{\epsilon}$ (Theorem 3) **then**
        // cloud side
        retrieve k documents from the indices
    else
        retrieve k documents from the k-out-of-k' OT protocol

A More Details about RemoteRAG

The detailed steps of RemoteRAG are shown in Algorithms 1 and 2.

We did not emphasize the detail of the k-out-of-k' OT protocol in module 2(c) in the main part, as it is not the primary contribution of this paper. However, for privacy, efficiency, and accuracy, we provide the detail and analysis of the k-out-of-k' OT protocol used in our experiments below.

A.1 k-out-of-k' Oblivious Transfer Protocol

We implement the k-out-of-k' OT protocol based on Chou and Orlandi (2015). Suppose the indices of target messages are $S = \{s_1, \cdots, s_k\}$, the protocol is as follows:

(1) The cloud and the user share a hash function Hash, a base number g and a prime modulus p.
(2) The cloud selects a random number a, computes $A = g^a \bmod p$, and sends it to the user.
(3) The user computes $B_i = A^{c_i} \cdot g^{b_i} \bmod p, i \in [1, k']$, where b_i are random numbers and $c_i = \begin{cases} 0, i \in S \\ 1, i \notin S \end{cases}$, and sends them to the cloud.
(4) The cloud constructs k' secret keys $\text{Key}_i = \text{Hash}({B_i}^a \bmod p), i \in [1, k']$, uses them to encrypt messages $[\![m_i]\!] = \text{Enc}(m_i, \text{Key}_i), i \in [1, k']$, and sends these encrypted messages to the user.
(5) The user constructs k secret keys $\text{Key}_{s_j} = \text{Hash}(A^{b_{s_j}} \bmod p), s_j \in S$, and can only decrypt target k messages $m_{s_j} = \text{Dec}([\![m_{s_j}]\!], \text{Key}_{s_j})$.

Correctness. The objective is to ensure that keys used for encrypting and decrypting target messages are consistent between both sides, while keys corresponding to other messages remain inconsistent.

▹ For $i = s_j \in S, c_i = 0$, the calculation of the key for m_i on the cloud side is ${B_i}^a \equiv (A^{c_i} \cdot g^{b_i})^a \equiv g^{ab_i} \bmod p$. Conversely, the calculation of the key for m_{s_j} on the user side is $A^{b_{s_j}} \equiv g^{ab_{s_j}} \bmod p$. This consistency in the calculation of the key on both sides enables the user to decrypt m_{s_j}.

Table 8: An example of recovered queries from perturbed embeddings.

Perturbation r	Recovered Query
Original Query	▹ My name is Alice. I got a cough. What should I do?
0	▹ ...... It's your name, Alice. If you are coughed, take some pills and
0.03	▹ ... You're a murmur.... You should take care of yourself. You collect isolated pieces of hair and
0.05	▹ read or take care of a tale of aluella. Aluellas are commonly spread and people have short patches of pneumonia, hair extensions or
0.07	▹ people suddenly get seriously affected, humans are allowed to stay off the walls and take care of a naturally occurring dementia. Alhaha viruses spread and
0.1	▹ Hair extensions of leucine A hair extensions of leucine are generally isolated or randomly formed bands of people who have long lived illnesses and take antibiotic or

▹ For $i \notin S, c_i = 1$, the calculation of the key for m_i on the cloud side is ${B_i}^a \equiv (A^{c_i} \cdot g^{b_i})^a \equiv g^{a(a+b_i)} \bmod p$. In contrast, the calculation of the key for m_i on the user side is still $A^{b_i} \equiv g^{ab_i} \bmod p$, if the user insists on generating a key. This inconsistency in the calculation of the key on both sides prevents the user from decrypting m_{s_i}.

Security analysis. The cloud receives $B_i = g^{ac_i+b_i} \bmod p, i \in [1, k']$. Since b_i is a random number generated by the user, the cloud cannot derive whether $c_i = 0$ or not by the given B_i. Therefore, the cloud has no idea of which indices the user chooses.

Communication analysis. There are 1.5 rounds of communication. First, the cloud sends a random number A, occupying β units. Second, the user sends $B_i, i \in [1, k']$ to the cloud, occupying $k'\beta$ units. Third, the user receives encrypted messages (documents) from the cloud, occupying $k'\eta$ units.

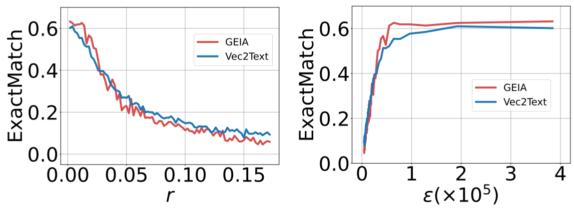
B Supplementary Results

B.1 Privacy Study

B.1.1 Query Reconstruction

To intuitively understand how the perturbation affects the recovery of the semantic meaning of queries, we provide an example below: The original query "My name is Alice. I got a cough. What should I do?" contains two main privacy information: the name "Alice" and her disease "cough". We add different sizes of perturbations to the query embedding, and apply Vec2Text to recover the query from perturbed embeddings. The results are listed in Table 8.

Without any perturbation, the recovered query contains both the name "Alice" and the disease



(a) ExactMatch-r (b) ExactMatch-ϵ

Figure 6: The ExactMatch metric corresponding to the perturbation r and the privacy budget ϵ.

"cough". When the size of perturbation is 0.03 or 0.05, we can still see "take care", which has a small chance to indicate that the user may describe something about health. However, the name "Alice" and the disease "cough" cannot be recovered from the attack. When the size of the perturbation is larger than 0.07, the recovered query seems to be entirely different from the original query.

B.1.2 ExactMatch

In Section 5.2, we use SacreBLEU to measure the difference between the original query and the recovered query from the sentence level. To provide a more fine-grained study, we leverage another metric called ExactMatch, which is also used by Morris et al. (2023) to measure the percentage of the recovered words that perfectly match the original query from the word level.

The results in Figure 6 show a trend similar to Figure 4. As the perturbation increases, the word in the original query becomes harder to reappear in the recovered query. The example in Table 8 also demonstrates the same situation. Therefore, we believe that keywords in the original query can be properly protected with (n, ϵ)-DistanceDP.

Table 9: Different data distributions do not affect the correctness of RemoteRAG.

	Datasets	k	r	Embedding Model
	NQ/TQA/MS	5/10/15/20	0.03/0.05/0.07/0.1	MiniLM/MPNet/T5/OpenAI-1/OpenAI-2
Recall	100%	100%	100%	100%

Table 10: In large-scale deployment, only the computation cost of module 1 increases.

Cost (N)	Module 1	Module 2(a)	Module 2(b)	Module 2(c)
Comp. (10^5) (s)	0.010	0.66	0.00014	0.014
Comp. (10^7) (s)	0.035 (×3.500)	0.66 (×1.000)	0.00012 (×0.857)	0.015 (×1.071)
Comm. (10^5) (KB)	6.18	38.44	2.04	63.63
Comm. (10^7) (KB)	6.20 (×1.002)	38.79 (×1.009)	1.92 (×0.941)	63.15 (×0.992)

B.2 Accuracy Study

B.2.1 Effect of Data Distribution

Apart from different dataset sizes, we also conduct experiments on the three datasets NQ, TQA and MS in full size, to inspect the effect of different data distributions. From the results in Table 9, we can still achieve 100% recall under different settings, demonstrating the correctness of RemoteRAG.

B.3 Efficiency Study

B.3.1 Large-Scale Deployment

For practical usage, we further increase the total number of RAG documents to 10^7 to see the costs of RemoteRAG in large-scale RAG service deployment. Table 10 summarizes computation and communication costs of each module in RemoteRAG. We find that compared to 10^5 RAG documents, only the computation cost in module 1 increases, while other costs remain nearly the same.

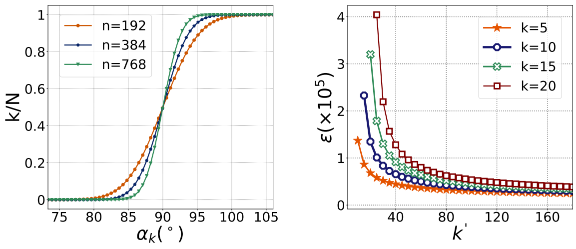
▹ In module 1, we retrieve top k' documents related to the perturbed embedding from N RAG documents, in which the cost of computation of distances is related to N. The communication cost of module 1 should remain the same since the only transmission is the perturbed embedding.

▹ In module 2, all operations are performed based on the k' documents retrieved from module 1. Due to unrelated to N, the costs should have no change.

However, even though the computation cost in module 1 increases, the total response time is still very short and acceptable.

B.4 Simulations

We use several simulation experiments to explore some details in RemoteRAG.



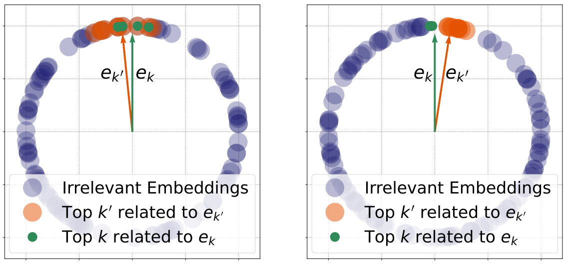
(a) k/N-α_k (b) ϵ-k'

Figure 7: Relationships among hyperparameters.

B.4.1 Relationships among Hyperparameters

$\boldsymbol{k/N}$**-**$\boldsymbol{\alpha_k}$**.** We first plot the equation of Lemma 1. From Figure 7(a), the result of k/N increases sharply as α_k approaches 90°. Additionally, when n grows larger, the increase is even steeper. This phenomenon is characteristic of high-dimensional space, where random vectors on the surface of the unit n-sphere tend to be almost perpendicular. Consequently, a relatively small change in α_k results in a significant change in k/N, meaning that the perturbation greatly impacts k', highlighting the importance of selecting a proper privacy budget.

$\boldsymbol{\epsilon}$**-**$\boldsymbol{k'}$**.** We discuss in Section 3.2 that in practice, apart from initially setting the privacy budget, we can also choose k' first and then compute the corresponding privacy budget. From Figure 7(b), we observe that when $k' < 50$, the change in ϵ is relatively large, which corresponds to a small perturbation and high attack performance according to Figure 4. The user should avoid considering k' as well as the corresponding privacy budget in this range, since the protection is too weak, as shown in Figure 4. To avoid excessive computation and communication costs, an appropriate choice of k'



(a) Visualization of top k. (b) One rare exception.

Figure 8: Illustrations in 2-dimensional space.

would be within the range of $[100, 200]$. Another observation is that a larger ϵ is required to preserve the same value of k' for a larger value of k. This can be explained by the fact that, as k increases, the number of possible embeddings with the same top k documents also increases. Therefore, the same value of k' implies a looser privacy requirement, which is reflected by a larger privacy budget ϵ.

B.4.2 2-Dimensional Simulations

The following 2-dimensional simulations are based on some artificial data for ease of understanding.

Visualization of Top k. To provide a clear view of how RemoteRAG works, we give a visualization of the relationship between the embeddings of the top k documents and the selected k' documents in 2-dimensional space. The plot in Figure 8(a) clearly shows that the top k documents are included in the set of k' documents, demonstrating the correctness of RemoteRAG.

Rare exceptions. Although we have not experienced any loss in retrieving documents, we acknowledge that in some rare exceptions, there might be a chance of RemoteRAG failing to preserve the top k documents. We provide one such exception here. As illustrated in Figure 8(b), the top k documents related to the query embedding are all located on the left side of the query embedding at the same angle α_k. The perturbed embedding is positioned on the right side of the query embedding at angle $\Delta\alpha_k$. If there are k' documents located exactly within the range of angles $[\alpha_k, \alpha_k + 2\Delta\alpha_k]$, then the top k' documents related to $e_{k'}$ would not include the top k documents related to e_k. However, exceptions like this only occur when the distribution of document embeddings is extremely non-uniform, a rare scenario in high-dimensional space. As a matter of fact, we have not encountered such an exception in our accuracy study (Section 5.3 and Appendix B.2).

C Proofs

Lemma 2 (Repeated from Lemma 1)**.** Assume that there are N embeddings uniformly distributed on the surface of the n-dimensional unit sphere. Let α_k be the polar angle of the surface area formed by top k embeddings related to any given embedding. Then, k and α_k satisfy the following relationship:

$$k = N \cdot \frac{\Omega_{n-1}(\pi)}{\Omega_n(\pi)} \cdot \int_0^{\alpha_k} \sin^{n-2}\theta \, d\theta$$

where $\Omega_n(\pi) = \frac{2\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2})}$ represents the surface area of the unit n-sphere.

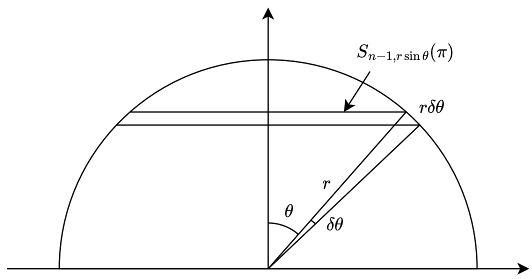


Figure 9: Illustration of the proof of Lemma 2.

Proof. Define $S_{n,r}(\alpha) = \Omega_n(\alpha)r^{n-1}$ as the surface area of the spherical sector with a polar angle $\alpha \in [0, \pi]$ in the n-sphere with radius r, where $\Omega_n(\alpha)$ represents the corresponding surface area in the unit n-sphere. Then, referring to Figure 9, we have

$$\begin{aligned} S_{n,r}(\alpha) &= \int_0^{\alpha} S_{n-1,r\sin\theta}(\pi) r \, d\theta \\ &= \int_0^{\alpha} \Omega_{n-1}(\pi)[r\sin\theta]^{n-2} r \, d\theta \\ &= r^{n-1} \int_0^{\alpha} \Omega_{n-1}(\pi) \sin^{n-2}\theta \, d\theta \end{aligned}$$

Comparing to the definition of $S_{n,r}(\alpha)$, it is straightforward to derive that

$$\Omega_n(\alpha) = \int_0^{\alpha} \Omega_{n-1}(\pi) \sin^{n-2}\theta \, d\theta$$

Assuming the embeddings are uniformly distributed on the surface,

$$\frac{N}{\Omega_n(\pi)} = \frac{k}{\Omega_n(\alpha_k)}$$

Therefore,

$$k = N \cdot \frac{\Omega_n(\alpha_k)}{\Omega_n(\pi)} = N \cdot \frac{\Omega_{n-1}(\pi)}{\Omega_n(\pi)} \cdot \int_0^{\alpha_k} \sin^{n-2}\theta \, d\theta$$

□

Theorem 4 (Repeated from Theorem 1)**.** Under the conditions specified in Lemma 2, given two embeddings e_k and $e_{k'}$ with the perturbed angle $\Delta\alpha_k$, to ensure that top k' embeddings related to $e_{k'}$ include top k embeddings related to e_k, k' and k satisfy the following relationship:

$$\Delta k = k' - k = N \cdot \frac{\Omega_{n-1}(\pi)}{\Omega_n(\pi)} \cdot \int_{\alpha_k}^{\alpha_{k'}} \sin^{n-2}\theta \, \mathrm{d}\theta$$

where $\alpha_{k'} = \alpha_k + \Delta\alpha_k$.

Proof. From Figure 3(b), we observe that $\alpha_{k'} = \alpha_k + \Delta\alpha_k$. This property can be readily extended to n-dimensional space. And from Lemma 2,

$$k = N \cdot \frac{\Omega_{n-1}(\pi)}{\Omega_n(\pi)} \cdot \int_0^{\alpha_k} \sin^{n-2}\theta \, \mathrm{d}\theta$$
$$k' = N \cdot \frac{\Omega_{n-1}(\pi)}{\Omega_n(\pi)} \cdot \int_0^{\alpha_{k'}} \sin^{n-2}\theta \, \mathrm{d}\theta$$

Thus,

$$\Delta k = k' - k = N \cdot \frac{\Omega_{n-1}(\pi)}{\Omega_n(\pi)} \cdot \int_{\alpha_k}^{\alpha_{k'}} \sin^{n-2}\theta \, \mathrm{d}\theta$$

□

Theorem 5 (Repeated from Theorem 2)**.** Given two normalized embeddings e_a and e_b of the same dimension, L2 distance and cosine distance have the following relationship:

$$d_{l2}(e_a, e_b) = \sqrt{2d_{\cos}(e_a, e_b)}$$

Proof. From Definition 2,

$$\begin{aligned} d_{l2}^2(e_a, e_b) &= \|e_a - e_b\|^2 = \sum_{i=1}^{n}(e_{ai} - e_{bi})^2 \\ &= \sum_{i=1}^{n} e_{ai}^2 + \sum_{i=1}^{n} e_{bi}^2 - \sum_{i=1}^{n} 2e_{ai}e_{bi} \\ &= \|e_a\|^2 + \|e_b\|^2 - 2\sum_{i=1}^{n} e_{ai}e_{bi} \\ &= 1 + 1 - 2\langle e_a, e_b\rangle \\ &= 2d_{\cos}(e_a, e_b) \end{aligned}$$

□

Lemma 3. k points $p_1, \cdots, p_k$ are extracted from the uniform distribution on the surface of an n-dimensional sphere with radius r. Denote the mean of these points as $\overline{p}$. L2 distance d between $\overline{p}$ and the center of the sphere has the expected value

$$\mathbb{E}[d] = \frac{r}{\sqrt{k}}$$

Proof. We place the center of the sphere at the origin. Since $p_i = \{x_{i1}, \cdots, x_{in}\}, i \in [1, k]$ is extracted from the uniform distribution on the surface of an n-dimensional sphere with radius r, the coordinates can be contructed by two steps: generating $y_{ij} \sim \mathcal{N}(0, 1)$ and $x_{ij} = r \cdot \frac{y_{ij}}{\sqrt{\sum_{m=1}^{n} y_{im}^2}}, j \in [1, n]$.

Due to symmetry, $\mathbb{E}[x_{ij}] = 0$,

$$\begin{aligned} &\mathbb{E}\left[\frac{y_{i1}^2}{\sum_{m=1}^{n} y_{im}^2}\right] = \cdots = \mathbb{E}\left[\frac{y_{in}^2}{\sum_{m=1}^{n} y_{im}^2}\right] \\ =&\frac{1}{n} \cdot \sum_{j=1}^{n} \mathbb{E}\left[\frac{y_{ij}^2}{\sum_{m=1}^{n} y_{im}^2}\right] = \frac{1}{n} \cdot \mathbb{E}\left[\frac{\sum_{j=1}^{n} y_{ij}^2}{\sum_{m=1}^{n} y_{im}^2}\right] = \frac{1}{n} \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}(x_{ij}) &= \mathbb{E}[x_{ij}^2] - (\mathbb{E}[x_{ij}])^2 \\ &= \mathbb{E}\left[r^2 \cdot \frac{y_{ij}^2}{\sum_{m=1}^{n} y_{im}^2}\right] = \frac{r^2}{n} \end{aligned}$$

By the central limit theorem, each coordinate component $\overline{x_j}$ of $\overline{p}$ satisfies

$$\overline{x_j} = \frac{1}{k} \cdot \sum_{i=1}^{k} x_{ij} \sim \mathcal{N}\left(0, \frac{r^2}{kn}\right), \frac{\sqrt{kn}}{r} \cdot \overline{x_j} \sim \mathcal{N}(0, 1)$$

Notice that

$$\frac{kn}{r^2} \cdot d^2 = \frac{kn}{r^2} \cdot \sum_{j=1}^{n} \overline{x_j}^2 = \sum_{j=1}^{n}\left(\frac{\sqrt{kn}}{r} \cdot \overline{x_j}\right)^2 \sim \chi^2(n)$$

Thus, $\frac{kn}{r^2} \cdot \mathbb{E}\left[d^2\right] = n$ and $\mathbb{E}[d] = \frac{r}{\sqrt{k}}$. □

Theorem 6 (Repeated from Theorem 3)**.** Given a target query embedding e_k and the mean embedding $\overline{e}$ of top k relevant document embeddings, the mean angle ω between e_k and $\overline{e}$ satisfies

$$\tan\omega = \frac{\tan\alpha_k}{\sqrt{k}}$$

where α_k is calculated from Lemma 2.

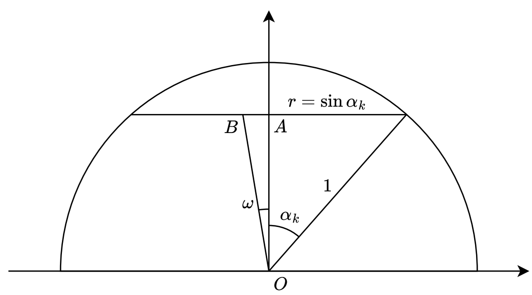


Figure 10: Illustration of the proof of Theorem 6.

Proof. Lemma 2 tells us that top k embeddings are within the polar angle α_k. When $n \gg 1$, we approximately believe that the angle they make with e_k is exactly α_k, which means k embeddings are uniformly distributed on the surface of an $(n-1)$-dimensional sphere with radius $\sin \alpha_k$.

Applying Lemma 3 and referring to Figure 10, $\mathbb{E}[AB] = \frac{\sin \alpha_k}{\sqrt{k}}$. Since $OA = \cos \alpha_k$,

$$\tan \omega = \frac{\mathbb{E}[AB]}{OA} = \frac{\sin \alpha_k}{\cos \alpha_k \sqrt{k}} = \frac{\tan \alpha_k}{\sqrt{k}}$$

□