A Lightweight and Communication-Efficient Protection Framework for RAG Retrieval via Heterogeneous Local Differential Privacy

Jiawei Hai
Hunan University, Changsha, Hunan Province, P.R. China
1183807201@qq.com

ABSTRACT

In order to solve the risk of privacy leakage in the Large Language Model (LLM) Search Augmented Generation (RAG), and the problems of "dimension disaster" and high communication cost faced by traditional local differential privacy (LDP) in high-dimensional space, this paper proposes a lightweight protection framework: reduced-dimension heterogeneous local differential privacy (DRH-LDP). In this framework, principal component analysis (PCA) is used to extract the core manifold of query vectors, and an innovative heterogeneous privacy budget allocation mechanism based on eigenvalue ratio is introduced to maximize the signal-to-noise ratio of core semantics. In addition, the fixed-scale 8-bit truncated quantization technique is used to compress the ciphertext and filter out the heavy-tailed noise. Experiments on real data sets show that under strict (, δ)-LDP constraints, the proposed method reduces the network traffic to 1/48 of the original size, and improves the retrieval recall rate by more than 5 times under privacy constraints. This research provides an efficient engineering paradigm for the safe deployment of RAG system in data-sensitive industries.

Keywords: Retrieval-Augmented Generation (RAG); Local Differential Privacy (LDP); High-dimensional Vector Retrieval; Heterogeneous Privacy Budget; Principal Component Analysis (PCA); Quantization Compression.

1. INTRODUCTION

Retrieval-augmented generation (RAG) technology is now being used in more and more different fields. Recent research has found some cool new directions: for example, combining RAG with clinical medicine textbooks can make intelligent question answering system more accurate and reliable, and also reduce the situation of artificial intelligence talking nonsense [1]; In addition, the text generation ability of LLMs and the retrieval enhancement function of RAG are used in file management [2]; In addition, scientists are also studying the potential of RAG architecture in battery material design, battery manufacturing, and battery management systems for electric vehicles and power grids [3].

However, now the user's query vectors can easily reveal privacy [4]. If we directly apply the traditional Local Differential Privacy (LDP) to these high-dimensional query vectors, we will often encounter the problem of "curse of dimensionality", which will lead to the poor retrieval recall [5]. Moreover, the commonly used methods of denoising or using complex cryptographic on the cloud now require a lot of network communication and computational overhead [6]. In this paper, we mainly want to solve the contradiction between "strict privacy protection" and "low-delay communication" in high-dimensional vector retrieval, and the goal is to solve the problem of excessive amplification of noise in traditional LDP and insufficient bandwidth between edge and cloud.

In order to solve these problems, this paper proposes and implements a lightweight protection framework of dimension-reduced heterogeneous local differential privacy (DRH-LDP). The system first uses Principal Component Analysis (PCA) to remove the redundant noise in the query vector, and then introduces a heterogeneous privacy budget allocation mechanism based on eigenvalue weight to replace the traditional uniform noise adding strategy. Before data transmission, we also added the technologies of Statistical Quantization and truncation to compress the ciphertext volume. Tests on real data sets show that this method effectively solves the performance bottleneck of traditional LDP in high-dimensional retrieval scenarios, not only greatly reduces the network overhead of edge-cloud interaction, but also provides a feasible technical reference for low-cost deployment of RAG systems in data-sensitive industries.

2. RELATED WORK

2.1 Overview of RAG Technology

Retrieval-augmented generation (RAG) is a hybrid generation architecture, which combines information retrieval technology with Large Language Models (LLMs) [7]. Its core idea is "retrieval first, then generation", which can be divided into three steps: retrieval, enhancement and generation. In the retrieval stage, after receiving the user's questions, the system will use semantic search technology to find relevant text fragments from massive external documents, which are not in the original training data of the model. In the enhancement stage, these retrieved documents will be put together with the original questions as prompt information to form input content with professional knowledge. This mechanism is equivalent to providing LLM with "temporary memory" or "reference materials", which greatly improves its ability to understand professional topics. Finally, in the generation stage, LLM will generate accurate, coherent and traceable answers according to the enhanced input content, which realizes the organic combination of pre-training knowledge and external knowledge [8].

2.2 Privacy-Preserving LLM Retrieval

With the increasing number of LLM, the problem of privacy leakage in RAG system has attracted great attention. Existing privacy-preserving retrieval schemes mainly rely on encryption technologies, such as Fully Homomorphic Encryption (FHE), Secure Multi-Party Computing (SMPC) or Trusted Execution Environments (TEE). Although these schemes can provide strict mathematical security, FHE will bring huge computational delay, and TEE is limited by hardware and may be attacked by side channels. In contrast, Differential Privacy (DP) is an ideal choice for edge devices with limited resources because of its low computational overhead.

2.3 Local Differential Privacy for High-Dimensional Vectors

Local Differential Privacy (LDP) allows users to add some noise before uploading data, so there is no need for a trusted central server. However, directly applying the traditional LDP method to high-dimensional and dense text vectors will encounter a serious "dimension disaster" problem-with the increase of dimensions, in order to meet the fixed privacy budget, the total amount of global noise will increase linearly, leading to a sharp decline in retrieval recall rate. At present, the research mainly tries to alleviate this problem by random sampling, hash bucket or feature space transformation before query [9]. However, these methods still can't keep fine-grained semantic features well, and they can't meet the needs of edge-cloud communication. The DRH-LDP mechanism proposed in this paper is to fill the research gap between high-dimensional semantic preservation and strict privacy protection.

3. METHODOLOGY AND SYSTEM DESIGN

In order to overcome the problems of "dimension disaster" and high communication overhead when using Local Differential Privacy (LDP) for high-dimensional dense vectors, this paper proposes a lightweight RAG retrieval protection architecture called Dimension-reduced heterogeneous local differential privacy (DRH-LDP).

The architecture adopts an edge-cloud decoupled design: the client side is responsible for query vector feature extraction, dimensionality reduction, heterogeneous noise injection, and quantization compression; the cloud-side vector database (e.g., FAISS [10]) is responsible only for receiving the processed low-dimensional ciphertext vectors and performing similarity searches. The overall system architecture and the decoupled edge-cloud workflow of the DRH-LDP framework are illustrated in Figure 1. The specific algorithmic workflow consists of the following three core steps:

3.1 Redundant Feature Stripping Based on Principal Component Analysis (PCA)

Query vectors generated by Large Language Models (LLMs) typically exhibit extremely high dimensionality (e.g., $d = 768$). However, the semantic information of natural language tends to concentrate within a low-dimensional manifold. If LDP noise is injected uniformly across all d dimensions, the noise in redundant "tail" dimensions will rapidly overwhelm core semantic features. Therefore, we introduce Principal Component Analysis (PCA) as a pre-transformation module on the client side.

First, a covariance matrix is computed offline using public corpora (e.g., Wikipedia). We extract the eigenvectors corresponding to the k largest eigenvalues to construct a projection matrix $W \in \mathbb{R}^{\mathbb{d}\times\mathbb{k}}$ (where $k \ll d$, setting $k = 64$). The target dimension k=64 was selected because the first 64 principal components retain the majority of the cumulative explained variance of the original semantic space. This configuration ensures the preservation of core topological features

for downstream retrieval, while limiting the communication overhead of a single query to exactly 64 Bytes under Int8 quantization constraints. When a user initiates a query at the client and generates an original high-dimensional vector$q \in \mathbb{R}^{\mathbb{d}}$, the system first projects it into the low-dimensional core manifold to obtain the reduced vector q':

$$q' = W^T q \tag{1}$$

Through this step, the system not only filters out redundant noise dimensions that contribute minimally to retrieval quality but also drastically reduces the number of dimensions consuming the privacy budget from d to k.

3.2 Heterogeneous Privacy Budget Allocation Based on Eigenvalues

Traditional Local Differential Privacy (LDP) mechanisms typically employ a uniform allocation strategy, where the total privacy budget ϵ is distributed indiscriminately across the k dimensions of the vector (i.e., the budget for a single dimension is ϵ/k). However, in the low-dimensional manifold space post-PCA projection, there are significant disparities in the semantic importance of each principal component. The variance information they contain follows a distinct diminishing trend (i.e., the variance of the first principal component is far greater than that of the k-th). Continuing to use a uniform allocation strategy in this low-dimensional space would lead to core features being drowned out by excessive noise.

To maximize query vector utility, this paper proposes a Heterogeneous Privacy Budget Allocation mechanism based on eigenvalue ratios. Let λ_i be the eigenvalue corresponding to the i-th principal component. The system dynamically segments the total privacy budget ϵ according to the weight of these eigenvalues within the total variance. The dedicated privacy budget ϵ_i for the i-th dimension is calculated as follows:

$$\epsilon_i = \epsilon \cdot \frac{\lambda_i}{\sum_{j=1}^{k} \lambda_j} \tag{2}$$

This allocation paradigm mathematically ensures that the primary components containing rich semantic information receive a larger privacy budget (thus incurring smaller perturbation noise), while subsequent dimensions with lower information entropy are allocated a smaller budget (incurring larger perturbation noise). By strictly adhering to the (ϵ, δ)-LDP privacy protection boundaries under sequential composition, this non-uniform adaptive noise injection strategy maximally preserves the core topological features of the vector, thereby significantly enhancing the system's final retrieval recall.

3.3 Gaussian Mechanism Noise Injection and Fixed-Scale Quantization

Following the heterogeneous allocation of the privacy budget, the system independently injects random noise satisfying the Gaussian mechanism into each dimension of the reduced vector. The noisy ciphertext dimension is expressed as:

$$\widetilde{q_i} = q_i' + \mathcal{N}(0, \sigma_i^2) \tag{3}$$

Where the standard deviation of the noise is inversely proportional to the privacy budget allocated to that dimension and directly proportional to the global sensitivity (Δf) of that dimension. The calculation formula is:

$$\sigma_i = \frac{\Delta f \sqrt{2\,ln(1.25/\delta)}}{\epsilon_i} \tag{4}$$

According to the Sequential Composition Theorem of differential privacy, the overall noise injection process strictly satisfies (ϵ, δ)-LDP privacy guarantees.

Then, in order to further reduce the network bandwidth occupied when transmitting encrypted vectors from edge devices to cloud, and to make the system more stable when searching in dense vector space, we add an 8-bit random quantization module with a fixed proportion after adding noise. This system will map and forcibly truncate the encrypted contents of Float32 into 8-bit discrete integers with a fixed scaling ratio. This quantization process can be expressed as:

$$\widehat{q_i} = \text{clip}\left(\left\lfloor \frac{\widetilde{q_i}}{s} \right\rfloor, -128, 127\right) \tag{5}$$

This design has two advantages: first, it can achieve a very high data compression rate; Secondly, because of the clipping property, it can effectively limit the unusually large outlier noise generated in tail dimensions (where the budget allocated is the least) and prevent it from interfering with the core semantics in L2 distance calculation in the cloud. Finally, based on the Post-processing Property of differential privacy, this quantization and truncation step is only used on the noisy

ciphertext to ensure that the original strict privacy boundary of the system is not affected.

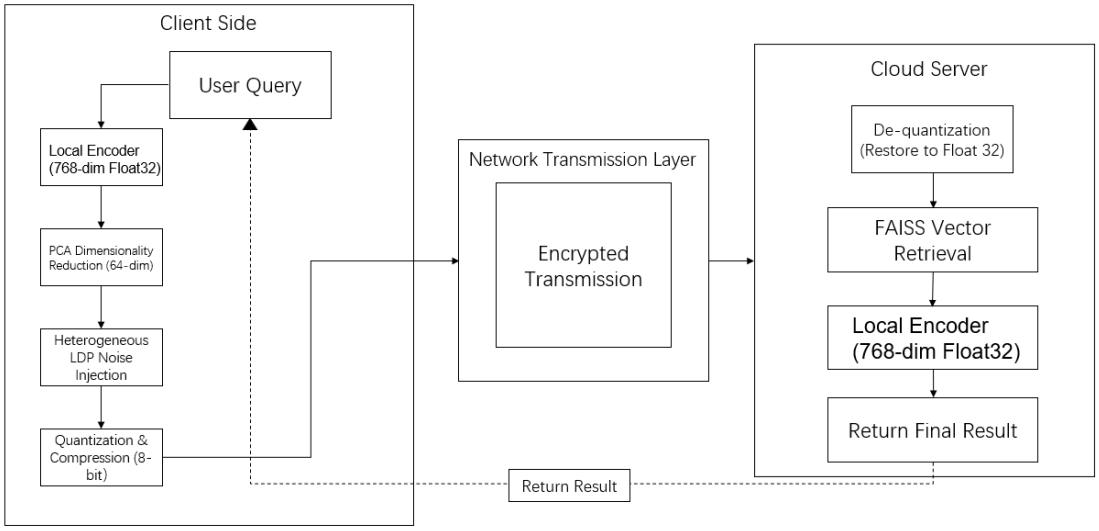

Figure 1. The overall architecture of the Dimension-Reduced Heterogeneous Local Differential Privacy (DRH-LDP) framework.

4. EXPERIMENTAL EVALUATION AND RESULTS ANALYSIS

In order to verify the usefulness of this method called dimension-reduced heterogeneous local differential privacy (DRH-LDP), this part shows a complete end-to-end evaluation on real data sets. Our analysis mainly focuses on the comprehensive performance of the whole system in terms of privacy protection intensity, retrieval recall rate and system overhead.

4.1 Experimental Setup and Evaluation Metrics

In order to test the performance of this system, our research used a data set called AG News to build a real full-process test environment. Specifically, we took out 2,000 written records from the training part of the data set to build a knowledge base; Then, 200 text records are randomly selected from the independent test part as the query set for retrieval test. In order to ensure that every experiment can be repeated, we set the global random seeds to 42.

In the step of feature extraction, we use a pre-training model called all-mpnet-base-v2 to turn sentences into high-dimensional codes. This model can turn all words into something called “dense vector”, and the initial output dimension is d=768. When searching, the system uses the precise Euclidean distance (also called L2 norm) to calculate the similarity, and uses the IndexFlatL2 indexing mechanism in FAISS library to do vector matching.

In order to eliminate the fluctuation of results caused by random interference and ensure the reliability of data, all performance data are averaged after 50 independent experiments. We chose the following baselines for comparison:

No-Privacy Bound: An idealized retrieval environment, where the query vector will not be disturbed, which represents the theoretical upper limit of the retrieval ability of the system.

Standard local differential privacy (Standard LDP): A traditional privacy protection scheme, which distributes the privacy budget evenly and directly adds Gaussian noise to the original 768-dimensional space.

The main evaluation indicators of this study include:

Hit Ratio (Top-10 Recall): It is used to evaluate the system’s ability to retain core semantic features and accurately find the target document under different privacy budgets.

System Communication Overhead: measured by Bytes/Query per query, it is used to evaluate the bandwidth efficiency of interaction between edge devices and the cloud.

4.2 Privacy-Utility Trade-off

The experiment first evaluates the fluctuations in retrieval recall under varying privacy budgets ϵ (as shown in Table 1).

Table 1. Comparison of Retrieval Recall and Communication Overhead across Different Privacy Mechanisms (Base Corpus: 2,000 records).

Retrieval Mechanism	Privacy Budget (ϵ)	Core Transmission Dimensions	Top-10 Recall (%)	Communication Overhead (Bytes/Query)
No-Privacy Baseline (Ideal)	N/A	768	100.0%	3072 (Float32)
Standard-LDP	1.0	768	16.4%	3072 (Float32)
DRH-LDP (Ours)	1.0	64	83.5%	64 (Int8)
Standard-LDP	2.0	768	55.4%	3072 (Float32)
DRH-LDP (Ours)	2.0	64	91.9%	64 (Int8)
Standard-LDP	4.0	768	87.0%	3072 (Float32)
DRH-LDP (Ours)	4.0	64	94.0%	64 (Int8)

When using the old Standard-LDP method, if we require high privacy protection (for example=1.0), the system will encounter a very troublesome “dimension disaster”. Because in the 768-dimensional high-dimensional space, noise will be piled up in a mess, and we have to distribute the limited privacy budget to each dimension equally, which leads to the query vectors being pushed far away from their original semantic groups. As a result, the recall rate of Top-10 suddenly dropped to only 16.4%, making the whole retrieval system almost useless.

On the contrary, the DRH-LDP mechanism proposed in this paper first compresses the feature space into a 64-dimensional low-dimensional space with PCA, and also adds an algorithm to dynamically allocate the privacy budget according to the proportion of the feature values. This method greatly reduces the noise interference in those unimportant tail dimensions. Under the same strict conditions (=1.0), the recall rate of DRH-LDP can reach 83.5%, which is more than five times higher than that of the traditional scheme. This shows that our new mechanism is much better at solving the performance bottleneck caused by high dimensions.

4.3 Analysis of System Communication Overhead and Truncation Protection

In addition to the breakthrough in retrieval practicability, our framework is also excellent in solving the bandwidth bottleneck of edge-cloud interaction. The traditional RAG retrieval scheme will directly transmit the high-dimensional Float32 encryption vector, and every query will consume 3072 bytes of bandwidth.

After adding noise, this paper introduces an 8-bit random quantization module, which can convert low-dimensional ciphertext into Int8 integer format. As shown in Table 1, this method not only greatly reduces the transmission capacity of each query to 64 bytes, but also uses the forced clipping function of 8-bit quantization to suppress abnormal noise. Because the tail dimensions are only allocated a small budget for privacy protection, they are particularly vulnerable to extreme noise; Forced truncation can effectively prevent these extreme values from having partial similarity results when calculating the distance in the cloud by setting the upper and lower limits of the values, thus ensuring the stability of ciphertext retrieval.

4.4 Case Study and Qualitative Analysis

In order to let people see the advantages of DRH-LDP mechanism in protecting query intention and resisting semantic drift more intuitively, we extracted a RAG retrieval case from a large AG News corpus for qualitative comparative analysis.

When a user initiates a query: “What is the latest news about space exploration and rockets?” At that time, the No-Privacy Baseline model accurately captured the intention of “Aerospace” and retrieved the target document: “redesigning rockets: NASA space propulsion findings a new home ...”. However, under the traditional Standard-LDP mechanism (=1.0), the Gaussian noise injected out of order in all dimensions causes serious semantic drift. The system completely lost the core

topological features of “Space” and “Technology” and was pushed to irrelevant semantic clusters by noise, which led to the wrong retrieval of documents related to politics or sports.

In contrast, the system named DRH-LDP-which has been dimensionalized by PCA, added with various noises, and truncated by Int8-can still adjust our problem vector in the group of “Technology and Aerospace” accurately and successfully find the file we want. This example shows that although the strict (, δ)-LDP mechanism will definitely bring some random interference to the data location, the core meaning structure of the problem vector can be well preserved by reducing the dimension, removing unimportant features and suppressing outliers with truncation. These results clearly prove that our method is really useful in dealing with the problem of “dimension disaster”, and also show that the whole system has found a good balance between protecting privacy and keeping it easy to use.

5. CONCLUSION

In order to solve the risks of privacy leakage in large-scale language model (LLM) search augmented generation (RAG) and the “dimension disaster” and communication bottleneck problems encountered by traditional local differential privacy (LDP) in high-dimensional space, this paper proposes a lightweight protection framework: reduced-dimension heterogeneous local differential privacy (DRH-LDP).

This framework first uses principal component analysis (PCA) to remove redundant features in query vectors. Secondly, it no longer uses the traditional one-size-fits-all noise injection strategy, but establishes a heterogeneous privacy budget allocation mechanism based on eigenvalue weight, which can maximize the signal-to-noise ratio of core semantics. Finally, an 8-bit quantization module is added to compress the ciphertext, and the abnormal noise in the tail dimension is effectively suppressed by forcibly truncating the boundary.

Experimental results show that under the strict (, δ)-LDP privacy protection condition, our proposed scheme improves the retrieval recall rate by more than 5 times compared with the traditional method, and also compresses the traffic of each query from 3072 Bytes to 64 Bytes (48 times). This research not only alleviates the effect decline caused by privacy protection in high-dimensional data, but also greatly reduces the network burden of edge computing and cloud interaction, and provides practical engineering reference for the safe deployment of RAG system in sensitive data fields such as medical care and finance.

REFERENCES

[1] Ding, N., Song, Y., Shan, Z., Dong, X., & Yu, M. (2025). Exploration on the construction of an intelligent Q&A system for medical teaching assistance based on Retrieval-Augmented Generation (RAG) technology. China Medical Education Technology, 39(1), 1-5. https://doi.org/10.13566/j.cnki.cmet.cn61-1317/g4.202501001

[2] Liu, Y. (2025). Analysis of the application of “Large Language Models+ RAG” technology in archival work. Archives of China, (3), 64-65.

[3] Zhong, Y., Leng, Y., Chen, S., Li, P., Zou, Z., Liu, Y., & Wan, J. (2024). Battery acceleration research based on Large Language Model RAG architecture: Current status and prospects. Energy Storage Science and Technology, 13(9), 3214–3225. https://doi.org/10.19799/j.cnki.2095-4239.2024.0604

[4] Koga, S., et al. (2024). Privacy-preserving retrieval-augmented generation with differential privacy. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 1024-1035).

[5] Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Now Publishers.

[6] Du, M., Wang, Y., & Lu, Z. (2023). Sanitizing sentence embeddings for local differential privacy. J. Inf. Secur., 45, 1542-1555.

[7] Zhan, D., Fan, S., Meng, X., Li, Y., & He, G. (2025). Research and practice of university intelligent Q&A systems based on locally deployed large models and RAG technology. Journal of Shenzhen Polytechnic University, 24(4), 94-100. https://doi.org/10.13899/j.cnki.szpuxb.2025.04.014

[8] Zhu, Y., & Li, S. (2025). Research on medium-wave knowledge base based on RAG technology. Audio Engineering, 49(12), 160–164. https://doi.org/10.16311/j.audioe.2025.12.048

[9] Chen, L., & Zhang, Z. (2025). Transform before you query: A privacy-preserving approach for vector retrieval. In International Conference on Learning Representations (ICLR) (pp. 45-58).

[10] Johnson, J., Douze, M., & Jégou, H. (2024). Faiss: A library for efficient similarity search and clustering of dense vectors [Computer software]. https://github.com/facebookresearch/faiss