Black Sea Journal of Engineering and Science
doi: 10.34248/bsengineering.1971136

Open Access Journal
e-ISSN: 2619 – 8991

Research Article
Volume 9 - Issue 4: 2039-2054 / July 2026

PRIVACY-AWARE ADAPTIVE DIFFERENTIAL PRIVACY FOR SEMANTIC RETRIEVAL: A PII-AWARE DYNAMIC BUDGET ALLOCATION FRAMEWORK

Seçkin MANDACI[1]*, Yılmaz VURAL[2], Özgür Can TURNA[1]

[1]*Istanbul University-Cerrahpasa, Faculty of Engineering, Department Of Computer Engineering, 34320, Istanbul, Türkiye*
[2]*Westcliff University, College of Technology and Engineering, Irvine, CA 92612, USA*

Abstract: Retrieval-Augmented Generation (RAG) makes it possible to enhance output from Large Language Models (LLMs) with retrieval-based evidence; however, these systems can also introduce privacy vulnerabilities, and in particular, Membership Inference Attacks (MIAs). Specifically, Differential Privacy (DP) is a well-established methodology for protecting privacy; however, applying static perturbation strategies, typically used in low-dimensional spaces, will result in degradation of semantic utility due to the high dimensionality of the embedding spaces. To solve this privacy-utility problem, we present PADP, a sensitivity-aware perturbation framework inspired by differential privacy principles, which provides a plug-and-play, middleware framework that can be easily integrated into enterprise RAG pipelines without requiring costly computations for LLM fine-tuning and reconstruction of vector indices. PADP uses Named Entity Recognition (NER) based on DeBERTa-v3 to perform sensitivity assessments of documents at the document level on an offline basis, computing PII sensitivity scores for each document. These scores are then used to adaptively calibrate the application of Laplace perturbations to dense query embeddings during retrieval. PADP assigns more protection to the higher-risk content and less to the lower-risk content than static perturbation methods would typically do, thereby increasing the effectiveness of retrieval. PADP draws on concepts from Differential Privacy; nonetheless, it does not furnish formal database-level Differential Privacy guarantees under conventional adjacency definitions. Consequently, the proposed framework should be understood as an adaptive privacy-enhancement mechanism rather than as a formally guaranteed Differential Privacy mechanism. Through the evaluation of the performance of PADP across diverse datasets by way of retrieval effectiveness metrics and distinguishability analysis, its comparison with static embedding perturbation baseline methods affords a contextual statistical analysis which indicates that adaptive sensitivity aware perturbation preserves semantic retrievability while dampening membership based distinguishability signals under the conditions of the study. These findings position sensitivity-aware perturbation as a practical approach to enhancing privacy in Retrieval-Augmented Generation systems while preserving utility.

Keywords: Retrieval-augmented generation, Differential privacy, Membership inference attacks, Privacy–utility trade-off, Semantic retrieval, Sensitivity-aware perturbation

***Corresponding author:** Istanbul University-Cerrahpasa, Faculty of Engineering, Department Of Computer Engineering, 34320, Istanbul, Türkiye
E mail: seckinm@gmail.com (S. MANDACI)
Seçkin MANDACI https://orcid.org/0000-0002-5556-4087
Yılmaz VURAL https://orcid.org/0000-0002-2858-5448
Özgür Can TURNA https://orcid.org/0000-0001-5195-8727

Received: June 15, 2026
Accepted: July 09, 2026
Published: July 15, 2026

Cite as: Mandaci, S., Vural, Y., & Turna, Ö. C. (2026). Privacy-aware adaptive differential privacy for semantic retrieval: a pii-aware dynamic budget allocation framework. *Black Sea Journal of Engineering and Science*, 9(4), 2039-2054.

1. Introduction

Large Language Models (LLMs) possess extensive capabilities that have changed how we work with language. They have been applied to many tasks related to natural language processing; for example: answering questions, producing summaries or assisting with conversations (Feng et al., 2024; Yao et al., 2024). Unfortunately, despite their many successes, LLMs are limited in their ability to incorporate new, dynamic knowledge sources into their decision-making processes. Specifically, LLMs utilize static parametric knowledge that they have learned through training. As a result of this, they may produce false or incorrect statements based upon new information sources that they have not been trained upon. A potential solution for addressing some of the challenges LLMs face as a result of their static nature is to leverage Retrieval-Augmented Generation (RAG). RAG has quickly become one of the most common methods for augmenting LLMs by providing access to external knowledge repositories via diverse types of databases when generating outputs from LLM Models. By incorporating information from various vector databases in real-time, models using RAG can greatly increase their factual reliability, as well as enhance their ability to generalize to new domains (Gade et al., 2026; Li et al., 2026).

Nonetheless, adding external retrieval functionality introduces new issues related to both privacy and security, including vulnerabilities associated with document ingestion and the broader RAG attack surface

This work is licensed (CC BY-NC 4.0) under Creative Commons Attribution 4.0 International License

(Castagnaro et al., 2025; Ward and Harguess, 2025; Ferrag et al., 2026). In contrast with many traditional language models, which typically exhibit vulnerability due to how much training data they remember, retrieval-augmented generation systems rely on external corpora that may hold sensitive corporate documents, proprietary information, or personally identifiable information. This means that the attack surface of these systems is far larger than the attack surface of traditional language model systems, greatly increasing the opportunities for those with nefarious intentions to engage in either membership inference attacks or attempts at direct data leakage (Anderson et al., 2025; Moia et al., 2026). Previous research has demonstrated how those with malicious intent could leverage both retrieval behavior and generation output to determine if certain documents exist in the knowledge base underlying the system, especially within domains with highly sensitive data, such corporate records or clinical records (Li et al., 2025; Vonderhaar et al., 2025; Wang et al., 2025; Azzam et al., 2026).

Utilizing Differential Privacy (DP) as an established method for limiting accessible knowledge about people by incorporating numerous properly calibrated randomization processes into our data processing capability is a well-known strategy (Dwork et al., 2006).

More broadly, perturbation-based anonymization and utility-oriented privacy-preserving data-publishing methods have sought to reduce disclosure risks while limiting information loss (Vural and Aydos, 2017; Eyupoglu et al., 2018).

There is a large difficulty in putting the conventional methods for DP into usage with a dense retrieval mechanism. Most of the standard approaches rely on static privacy budget, which allows the application of the same perturbation strength on the same content, irrespective of how sensitive that content may be to an individual. With high-dimensional embedding spaces, static perturbation mechanisms can degrade the quality of semantic retrieval, due to dimensionality issues (Habernal, 2021; Habernal, 2022) associated with an increasing number of dimensions (i.e., increasing number of embedded dimensions for an item of content). To protect an individual's privacy, DP typically requires increasingly strong perturbation power as the number of embedding dimensions increases, thus unnecessarily compromising access to (i.e., retrievability of) content that may not contain sensitive data. Current adaptive implementations typically require retraining a language model, modification to the embedding representation, or converting the corpus to synthetic data; therefore, limiting the utility of the adaptive implementations as they relate to the current state of dynamic enterprise environments (Ge et al., 2026).

This research presents the Privacy-Aware Adaptive Differential Privacy (PADP) framework to overcome the operational and semantic constraints experienced by retrieval-augmented generated systems. Instead of applying uniform perturbation over the entire retrieval pipeline, PADP evaluates document-level sensitivity off-line by determining the presence and relative degree of Personally Identifiable Information (PII) in the document with a DeBERTa-v3 Named Entity Recognition (NER) model (Mainetti and Elia, 2025; Azzam et al., 2026). The entities identified through this process are assigned weighted risk measures based on the degree of sensitivity (Zeng et al., 2026), which provides a means to calculate normalized sensitivity scores for each individual document.

The precomputed sensitivity scores obtained from the previous retrieval step serve as a guide for adapatively calculating Laplacian perturbations that are directly applied to dense embeddings from the query. Therefore, PADP will apply more security to content that has a greater risk of violation of the privacy compared to content that has less risk of privacy violation and will be effective at retrieving content. Unlike many other methods in which the document collection would have to be extensively retrained or vectors would need to be repaired, PADP has been developed as a plug-and-play middle-ware that can be easily integrated into any existing RAG based ecosystem (e.g., LangChain and LlamaIndex) without having to modify any of the underlying LLM's.

It should be noted that PADP is an operational privacy mechanism influenced by differential privacy, rather than a framework offering strict differential privacy guarantees on an end-to-end basis for all database records. Because adaptive perturbation strengths are based upon the sensitivity information derived from the database corpus, to date, rigorous privacy accounting using standard definitions of adjacency is still an open area of research. Nonetheless, the function of providing sensitivity-aware perturbations can be used as a pragmatic technique to optimise privacy-utility trade-offs in practice.

The following is a summary of the contributions gone over in depth throughout this study:

Retrieval-Time Sensitive-Aware Middleware for Privacy: PADP is a middleware that requires no retraining or vector index reconstruction on the Large Language Model (LLM) to integrate with existing Retrieval-Augmented Generation (RAG) systems as a plug-and-play middleware solution.

Document-Level Sensitivity Assessment for Offline Assessment: This paper presents a new framework for assessing the sensitivity level of documents using DeBERTa-v3 based PII detection aggregated with a weighted scheme to determine privacy based upon contextual sensitivity of documents.

Dynamic Adaptive Perturbation of Embedded Representations for Sensitivity: PADP applies Laplace perturbation by dynamically calibrating the amount (or magnitude) of Laplace perturbation according to the sensitivity associated with a given document, with the overall objective of achieving an appropriate balance

between privacy protection and retrieval effectiveness in heterogeneous repositories.

Analysis of Privacy-Utility Tradeoffs with Empirical Testing: Adaptive perturbation better protects the utility of retrieval than static perturbation methods provided that the same experimental conditions are maintained, while also decreasing the level of distinguishability between retrievals in relation to the associated risks of Membership Inference Attack (MIA) occurring. Examination of Limitations of Adoption of Adaptive Privacy Allocation Method: Limitations exist with adopting adaptive privacy allocation practices, including the lack of formal Database-Level Differential Privacy (DP) guarantees and the need for more rigorous evaluation of adaptive privacy allocation practices against existing/more progressive patterns of MIA.

1.1. Related Works

1.1.1. Security and privacy risks in RAG systems

The application of Retrieval-Augmented Generation (RAG) systems which integrate external knowledge repositories with Large Language Models (LLMs) has led to improvements in factual correctness and generalization capabilities of these models. This integration also creates additional security and privacy risks, including prompt-injection attacks that may originate from user queries or retrieved documents (Geng et al., 2026). The security and privacy risks associated with LLM systems include training-data memorization, membership inference, and data extraction, whereas RAG systems introduce additional risks through the external retrieval collections used to support their functionality (Yan et al., 2025). This means that RAG systems will also have additional attack surfaces that allow them to be exploited through prompt injection and manipulation of retrieval data and/or exposed to data leakage through the third-party data provided in the external retrieval sets. As demonstrated in recent literature reviews and systematic literature reviews regarding RAG systems, the use of external vector databases creates an increase in the number of potential attack vectors exposing the RAG pipelines to risks associated with prompt injection, manipulation of retrieval data, and data leakage risks; however, these increased attack vectors to RAG pipelines are particularly important in high-stakes, high-security domains, specifically healthcare and enterprise applications where the unauthorized disclosure of any retrieved information could have devastating impacts (Vonderhaar et al., 2025; Azzam et al., 2026).

1.1.2. Membership inference attacks in RAG

Previously, Membership Inference Attacks (MIAs) have been focused on finding out if a certain record(s) were used in creating the model (Maini et al., 2024; Chen et al., 2025; Puerto et al., 2025). However, more recently it has been found by researchers there are new ways to do MIAs against retrieval-based architectures that target the retrieval corpus (Anderson et al., 2025; Li et al., 2025). We see literature has currently proposed several attack strategies for doing an MIA. Interrogation-based attacks use carefully constructed natural language queries to determine if the target document(s) exist. Likelihood-based attacks utilize various conditional probability differences and consistency in the way they retrieve information to estimate an MIA (Xie et al., 2024; Xin et al., 2024; Qi et al., 2025). More complex attacks, include Differential Calibration MIA (DCMI), Difficulty-Calibrated MIA (DC-MIA), BudgetLeak, and black-box style attacks can be used to demonstrate RAG's vulnerability based on realistic work operations assumptions (Gao et al., 2025; Wang et al., 2025). All the research and studies support the belief that protecting against MIAs of retrieval corpuses is an outstanding and still-to-be-solved-problem.

1.1.3. Differential privacy in dense retrieval

Differential Privacy (DP) serves as an established theoretical basis to regulate the amount of information disclosed through random perturbation techniques (Dwork et al., 2006; Pan et al., 2024). However, applying typical DP methods directly to the problems associated with dense retrieval systems creates numerous significant challenges. Standard methodologies use static perturbation methods without regard for the sensitivity of the data by applying identical amounts of privacy budget; in this case, static perturbation on high-dimensional embedding spaces can alter the semantic representation of items significantly, which in turn results in reduced retrieval performance because of the curse of dimensionality (Habernal, 2021; Habernal, 2022). Furthermore, as the embedding's dimensionality increases, so too does the requirement for stronger perturbation to preserve the privacy of each document, which could impede the utility of documents with little to no sensitivity or risk (Sanchez-Serrano et al., 2025; Wu et al., 2026). Despite several attempts to investigate selective and adaptive DP mechanisms for broader applications in the field of NLP (Wang et al., 2024), there currently exist difficulties in establishing an effective balance between privacy protection and retrieval utility, particularly when operating in a dynamic RAG framework.

1.1.4. Adaptive DP, PII detection, and the positioning of PADP

Recent studies have suggested multiple mechanisms including differentially private generation, to maintain privacy throughout various stages of the RAG pipeline (Grislain, 2025). Training stage techniques have been developed to maintain privacy through the application of differential privacy during model creation and the generation of synthetic datasets that help maintain privacy. Most of these techniques have the ability to significantly increase privacy; however, they can have a large computational cost (as they require significant computational resources and also may involve retraining models that contain useful information, as the underlying local data may change frequently).

Inference stage techniques, including locally private

perturbation of sensitive entities, have also started to gain considerable attention (He et al., 2025). For example, the PRW technique for decoding uses various techniques to perturb sensitive information when generating results while other queries can operate using standard cryptographic contracts to protect the privacy of the user's query (e.g., aPRW). PRESS uses embedding space manipulation techniques to reduce the risk of information leakage, while LPRAG uses a variety of perturbation techniques applied at the system level to perturb sensitive information located within the context of retrieved results. The design choices highlighted by these different privacy mechanisms provide interesting trade-offs along multiple measures of privacy confidence—including privacy assurance levels, computational costs associated with using various privacy mechanisms and the level of success of the retrieval process. There are still many open challenges to the privacy protection question in the literature. The vast majority of current techniques are based on static or historical perturbation methods, create large-scale infrastructure modifications, rely on synchronous detection mechanisms that could contribute to latency and have very little information published concerning the privacy vs. utility trade-off of the retrieval embedding.

1.1.5 Positioning of PADP

The new Privacy-Aware Adaptive Differential Privacy (PADP) framework was created to address these long-standing challenges; it provides enhanced security for content that has a higher risk of exposure by utilizing offline assessments of sensitivity at the document level and allowing for an adaptive alteration in data during retrieval, which provides continued effectiveness of retrieval of lower risk documents. PADP's "plug-and-play" ability allows it to be integrated into currently used RAG infrastructure where other solutions require retraining language models or rebuilding vector indices. The PADP approach is meant to be viewed as one of many privacy methods based on principles of differential privacy, an operating method, rather than a definitive solution to retrieval privacy. Therefore, it should serve as a mechanism to improve the privacy/utility trade-off given realistic constraints when deployed. PADP is different from the existing differential-privacy-based RAG in three main ways. First, it does not require any retraining of the underlying language model or re-building of the vector index. Second, rather than applying a uniform perturbation strategy to all queries or documents, it uses document-level PII sensitivity scores for adaptive privacy-budget allocation. Thirdly, PADP works as a retrieval-time middleware layer that can easily fit into an already existing enterprise RAG pipeline. Thus, this framework is more like a practical mechanism to enhance privacy with awareness of sensitivity rather than being completely substitutive for formally differentially private training or indexing procedures.

Further evaluation against greater levels of threats than currently exist and formal privacy analyses are necessary for future research. The remainder of this paper is organized as follows: Section 2 introduces the system's threat model and the algorithmic foundations of the PADP framework. Section 3 presents comprehensive experimental findings; Section 4 discusses the operational limitations of the adaptive mechanism, and Section 5 concludes the paper.

2. Materials and Methods

This section details the formal mathematical definition and operational mechanics of the proposed Privacy-Aware Adaptive Differential Privacy (PADP) framework. The methodology integrates a rigorous threat model, an un-intrusive plug-and-play architecture, offline context-aware local sensitivity indexing, online risk-proportionate embedding perturbation, and formal privacy composition guarantees.

2.1. Threat Model and Plug-and-Play System Architecture

To formalize the privacy vulnerabilities within the Retrieval-Augmented Generation (RAG) pipeline, we establish a strict black-box adversary model consistent with the standard Dolev-Yao security framework (Anderson et al., 2025; Moia et al., 2026). In this architecture, the adversary has zero operational visibility or structural access to the localized parameters of the underlying Large Language Model (LLM), the raw text sequences inside the external vector database, or the intermediate embedding layer tensors. The interaction plane is strictly bounded by an API interface where the adversary issues natural language queries q and observes the finalized, textually synthesized responses y. Despite these black-box boundaries, the adversary is assumed to execute advanced semantic reconnaissance. This includes the capacity to systematically probe the system via adaptive query generation and to expt token-level output distributions using relative conditional log-likelihood metrics (Li et al., 2025; Wang et al., 2025). The ultimate objective of the adversary is to execute a Membership Inference Attack (MIA), thereby determining whether a specific target document d is present within the retrieval corpus r by maximizing the posterior extraction probability P (M = 1 q, y, r). In contrast to existing defense mechanisms that introduce significant runtime overhead via model fine-tuning or full corpus synthesis T. Wang et al. (2024), PADP operates entirely as a non-intrusive middleware layer. It introduces no adjustments to the vector index topology or the core LLM parameters, enabling seamless, plug-and-play integration into production-grade enterprise RAG orchestration frameworks such as LangChain or LlamaIndex.

2.2. Offline Context-Aware Local Sensitivity Indexing

The core algorithmic contribution of PADP is its capability to map fine-grained privacy boundaries by measuring document-level Personally Identifiable Information (PII) density to establish local sensitivity upper bounds. To eliminate computational bottleneck and runtime latency during the online evaluation

window, this step is executed entirely offline during the document onboarding and indexing phase. Rather than utilizing legacy rule-based pipelines (e.g., standard regular expressions or vanilla spaCy tokenizers) which cause substantial false-positive rates in unstructured texts, PADP leverages a highly accurate Transformer model optimized with a disentangled attention mechanism (DeBERTa-v3-pii-masker) (Azzam et al., 2026; Mainetti and Elia, 2025; Szawerna et al., 2024).

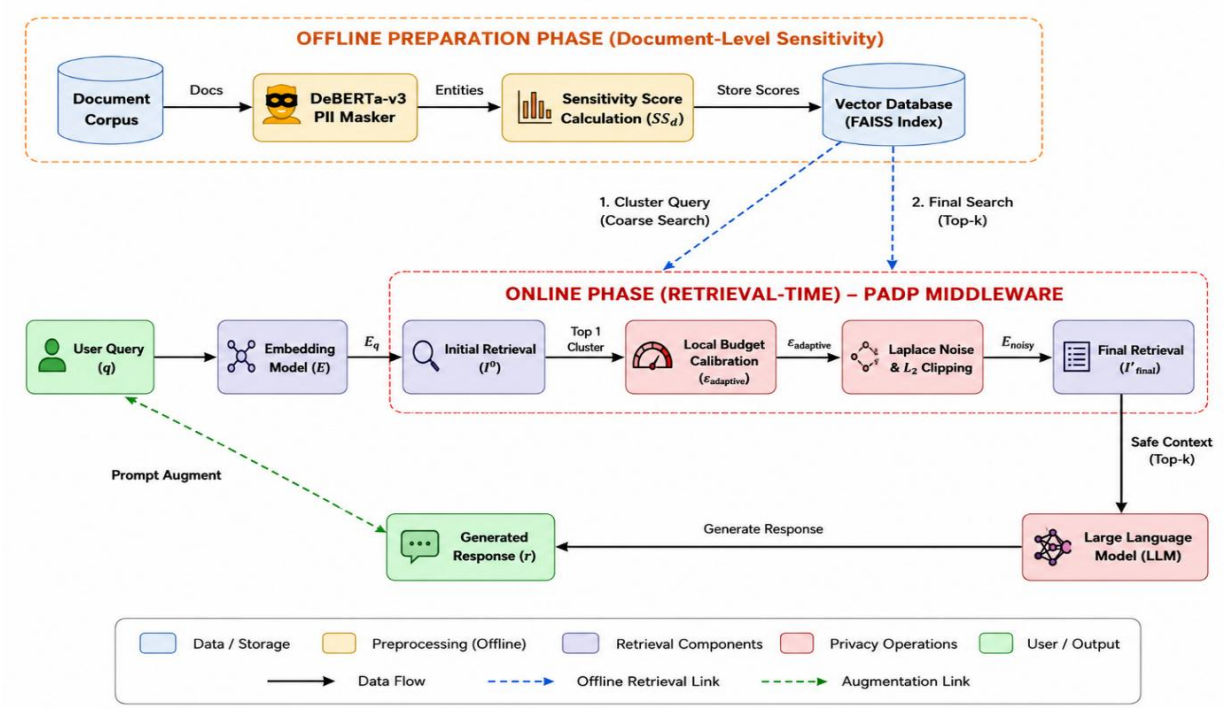

Figure 1. The structured architectural blueprint of PADP, featuring a balanced, closed-loop visualization of offline LDP sensitivity bucket indexing and online perturbation tracking mechanics.

The architecture parses each text entry d within the data store to locate precise PII categories, matching each entity type to an asymmetric risk weight wi scaled according to its relative vulnerability severity. In our setup, these risk scales are parameterized as follows: high-risk alphanumeric identifiers (e.g., SSN, passports tax records) are assigned wi = 5.0; communication vertices (e.g., email, telephone numbers, physical addresses) are assigned wi = 3.0; core biographical identifiers (e.g., individual names, organizations) are assigned wi = 2.0; and basic temporal indices (e.g., dates, ages) are as-signed wi = 1.0. Let d represent the set of extracted PII entity categories in document d, and let ct(d) denote the frequency count of entity type t. The total accumulated risk mass $N_{weighted}$ is defined as (equation 1):

$$N_{\text{weighted}} = \sum_{t \in \varepsilon_d} w_t \cdot c_t(d) \#(1) \quad (1)$$

To ensure statistical comparability across documents of heterogeneous token lengths and accurately represent sensitive density, the context-aware Sensitivity Score (SSd) is formulated as a smooth bounding function (equation 2):

$$SS_d = \frac{N_{\text{weighted}}}{N_{\text{weighted}} + T_{\text{threshold}}} \quad (2)$$

where Threshold models an empirical scaling constant set to 10.0. This formulation guarantees that SSd [0, 1), structurally locking each document's local sensitivity boundary inside the Local Differential Privacy (LDP) directory at the time of database ingestion.

2.3. Online Risk-Proportionate Budget Allocation

Enforcing a uniform, static privacy budget across a structurally diverse knowledge base severely degrades the semantic utility and retrieval fidelity of non-sensitive records Habernal (2021). To bypass this data-dependent leakage where the perturbation scale itself leaks details of the target query, PADP dynamically allocates the localized budget ϵadaptive at retrieval-time based on the precomputed offline local sensitivity upper bound (SSd) of the cluster index rather than real-time content inspection Wu (2025) (equation 3):

$$\epsilon_{\text{adaptive}} = \frac{\epsilon_{\text{base}}}{1 + \lambda \cdot SS_d} \quad (3)$$

where the baseline global budget constant ϵbase is set to 10.0 and the scaling hyperparameter λ is set to 10.0. This mathematical structure ensures that documents devoid of PII (SSd 0) maintain an elevated privacy budget, preserving semantic retrievability. Conversely, documents mapping to high-risk profiles (SSd 1) trigger an immediate constriction of the budget scale, enforcing aggressive noise bounds. Unlike localized token perturbation schemes He et al. (2025), these runtime

calibration shifts are calculated against static, pre-cataloged offline thresholds, The adaptive perturbation strategy is intended to improve the privacy–utility trade-off by allocating stronger perturbation to higher-risk content while preserving retrieval effectiveness for lower-risk documents.

2.4 High-Dimensional Embedding Manifold Perturbation and L_2 Clipping

Because semantic lookup within production RAG workflows is conducted over continuous dense vector representations (Ye et al., 2025b), PADP injects its risk-proportionate perturbation directly into the embedding space during the online retrieval sequence. Let $V_q \in R^{384}$ denote the dense localized vector representation of the incoming query. Prior to finalizing candidate document ranking, the middleware perturbs the state representation (equation 4):

$$V_{\text{noisy}} = V_q + \eta \tag{4}$$

where each individual j-th dimension of the multidimensional noise vector η is drawn independently from a zero-mean Laplace distribution scaled inversely by the adaptive budget (equation 5):

$$\eta_j \sim \text{Laplace}\left(0, \frac{\Delta f}{\epsilon_{\text{adaptive}}}\right) \tag{5}$$

Here, Δf models the global $\ell 1$-sensitivity parameter of the underlying embedding function, fixed empirically at 0.15. In high-dimensional vector spaces (d-dimensional space), injecting independent noise across each individual axis causes the total spatial deviation of the cumulative vector to expand asymptotically with the square root of the dimensions, ($O\sqrt{d}$) (Habernal, 2022; Wu et al., 2026b). To control this curse of dimensionality and decouple the localized differential privacy guarantee from the spatial size of the vector space, we enforce a strict spherical L2 clipping parameter (equation 6):

$$\|\eta\|_2 \leq C_{\text{clip}} \cdot \|V_q\|_2 \tag{6}$$

where the operational clipping threshold bound C_{clip} is fixed at 10.0 to maintain directional semantic stability. The clipping threshold was selected empirically to preserve directional stability and was not optimized theoretically.

2.5. Algorithmic Pipeline Workflow

The runtime execution sequence of the PADP framework within the enterprise retrieval workflow is formalized in Algorithm 1. This pipeline demonstrates the non-intrusive interception characteristics of the middleware.

Algorithm 1 PADP Online Inference-Time Pipeline

Input: Query q, Vector Corpus $\mathcal{C}$, Embedding Model ϕ, Global Sensitivity Δf, Base Budget $\epsilon_{\text{base}} = 10$, Clipping Constant $C_{\text{clip}} = 10$

Output: Secure Retrieved Document D_{final}, LLM Response y

$V_q \leftarrow \phi(q)$ {Generate original dense query embedding}
$\text{Bucket}_{\text{cand}} \leftarrow \text{FAISS.retrieve}(V_q, \mathcal{C}, k = 1)$ {Lookup the target sensitivity cluster}
$SS_d \leftarrow \text{Fetch\_Offline\_SS}(\text{Bucket}_{\text{cand}})$ {Retrieve precomputed cluster sensitivity map}
$\epsilon_{\text{adaptive}} \leftarrow \epsilon_{\text{base}}/(1 + 10 \cdot SS_d)$ {Calibrate risk-proportionate local budget}
Draw $\eta \sim \text{Laplace}\,(0, \Delta f/\epsilon_{\text{adaptive}})^{384}$ {Generate high-dimensional noise matrix}
if $\|\eta\|_2 > C_{\text{clip}} \cdot \|V_q\|_2$ **then**
  $\eta \leftarrow \eta \cdot \frac{C_{\text{clip}} \cdot \|V_q\|_2}{\|\eta\|_2}$ {Enforce L_2 clipping against dimension expansion}
end if
$V_{\text{noisy}} \leftarrow V_q + \eta$ {Perturb query state within embedding manifold}
$D_{\text{final}} \leftarrow \text{FAISS.retrieve}(V_{\text{noisy}}, \mathcal{C}, k = 1)$ {Execute finalized secure semantic search}
$y \leftarrow \text{LLM.generate}(q, D_{\text{final}})$ {Synthesize context-grounded secure response}
return D_{final}, y

Note: PADP provides practical privacy enhancement guided by offline sensitivity assessment and should not be interpreted as providing strict database-level differential privacy guarantees.

2.6 Privacy Interpretation and Evaluation Protocol

The Perturbed Adaptive Data Protection Protocol (PADP) employs Laplace-Calibrated Perturbations that are similar to techniques that are often referred to as 'differential privacy' mechanisms based upon prevailing differential privacy mechanisms and algorithms. Yet, the strength of adaptive perturbation depends upon the presence of (i) both the sacrifice the privacy of individual documents and (ii) the privacy of all documents during to the indexing corpus in offline mode to determine the sensitivity of documents. Hence, the intended interpretation of this current formulation does not provide a formal end-to-end database-level or a differential privacy guarantee, as expressed under the standards of adjacency.

Instead, the operative model of PADP should provide an operationally practical mechanism for increasing the usefulness of retrieving data from a privacy-preserving and data retrieval mechanism. In practice, higher levels of perturbation should be assigned to potentially riskier data, as determined by how similar the data is to each other when compared to the total data published publicly, while retaining the semantic utility of lower-risk documents.

The privacy accounting for adaptive retrieval mechanisms is still a significant opportunity for researchers to further understand and develop extremely

reliable and consistent privacy mechanisms that meet or exceed the level of privacy afforded during retrieval by the use of these two components.

The empirical testing of both the effectiveness of the retrieval operation and the distinguishability of the reader's data after retrieval, which are employed to describe the trade-offs between privacy and utility, includes the use of the Wilcoxon signed-rank test and the effect size estimate to permit the evaluation of multiple conditions using statistical methods, such as the Kolmogorov-Smirnov statistics. These tests will provide sufficient information to operationalize the level of privacy provided; however, they do not provide complete security from membership inference attacks that exist at this moment.

For interpretability purposes, the equivalent privacy indicator (εeq) is reported as a descriptive summary of observed adaptive perturbation strengths and should not be interpreted as a formally composed differential privacy budget.

The hyperparameters used in PADP were selected empirically to provide a stable privacy–utility balance across heterogeneous datasets. The scaling parameter λ governs the intensity of sensitivity-aware budget adaptation, where larger λ values amplify perturbation for documents with greater sensitivity scores. The sensitivity parameter Δf sets the baseline level of Laplace noise injected into the embedding representation. The clipping threshold C_clip is applied to avoid excessive distortion in high-dimensional embedding space and maintain semantic directionality. Lastly, T_threshold is introduced as a smoothing constant in the calculation of the sensitivity score to prevent unstable sensitivity estimates for documents with very low or very high PII counts. These parameters are not optimized theoretically but rather were selected to ensure consistent experimental behavior and interpretable comparisons across datasets.

2.7. Datasets

Table 1 summarizes the datasets used throughout the experimental evaluation. The selected datasets represent diverse textual domains with varying privacy requirements, ranging from general news classification to highly sensitive PII-rich synthetic records. This diversity allows the proposed PADP framework to be evaluated under heterogeneous privacy conditions and provides a realistic assessment of its robustness across different sensitivity levels.

3. Results

3.1. Overall Retrieval Performance

Results appear in Table 2 along with Figure 2, showing retrieval performance across conditions. Best average scores came from No-DP, a predictable result given its lack of privacy-induced noise. When privacy was preserved, PADP outperformed others under both disturbance settings. With Delta_f set to 0.05, Recall@1 reached 0.1296 using No-DP, dipped to 0.1120 with Static DP (epsilon = 5.0), then climbed to 0.1248 when PADP was applied. At a Delta_f of 0.15, the recorded figures stood at 0.1296, 0.0468, and 0.1136 in order. Though perturbations grew stronger, PADP maintained much of the original method's effectiveness - outperforming Static DP by a clear margin. Despite increasing noise levels, performance loss remained limited under PADP. The gap between methods widened as conditions became less favorable. Under higher disturbance, one strategy clearly held its ground better than the other.

Across each dataset, patterns point in one direction. When Delta_f was set to 0.15, PADP reached a Recall@1 of 0.140 on IMDB - followed by 0.114 on PubMed, then 0.126 on AG News. CNN/DM showed a lower value at 0.052, while SynthPII recorded 0.136 under the same method. On the flip side, using Static DP with epsilon at 5.0 led to just 0.078 on IMDB. Values dropped further: 0.056 (PubMed), 0.042 (AG News), 0.012 (CNN/DM), and 0.046 (SynthPII). Every number taken together suggests PADP pulled ahead every time.

3.2. PADP versus Static DP

Figure 3 along with Table 3 shows how PADP performs compared to Static DP at epsilon = 5.0. With Delta_f fixed at 0.15, gains in Recall@1 reached 0.062 on IMDB, followed by 0.058 on PubMed, then climbed to 0.084 on AG News, edged up 0.040 on CNN/DM, and peaked at 0.090 on SynthPII. Despite tighter perturbation settings, Under lower perturbation levels, improvements were generally small and not statistically significant. Larger gains emerged under stronger perturbation settings.- effect sizes ranged from moderate to strong according to Hedges' g. On CNN/DM, the jump stood out clearly because baseline performance with Static DP started so low. Yet when measured purely by magnitude, SynthPII showed the largest lift overall. Even at Delta_f = 0.05, PADP still outperformed Static DP on every dataset, but by smaller amounts. Because the perturbation was less intense, Static DP managed to keep more of its original effectiveness. What stands out here is that PADP shines most when privacy demands are toughest - exactly where standard fixed-budget approaches lose accuracy faster.

3.3. Privacy-utility trade-off

The graph labeled Figure 4 maps how privacy and usefulness relate, measured by 1/(1+epsilon_eq) for privacy and average Recall@1 for performance. Rather than balancing well, fixed Differential Privacy at epsilon levels of 0.5 and 1.0 gave high privacy yet weak results in finding correct matches. When epsilon rose to 5.0 under the same method, accuracy improved - though protection weakened noticeably. On the other hand, PADP maintained a steadier position between these two goals, especially when Delta_f was set to 0.15; here, it outperformed Static DP with epsilon=5.0 in recall without sacrificing as much privacy. What stands out is that PADP exhibited a favorable operating point among the evaluated conditions.

Table 1. .Dataset characteristics used in the experimental evaluation

Dataset	Domain	Approximate Samples	Sensitivity Level	Primary Privacy Concern
IMDB	Movie Reviews	50,000	Medium	Personal opinions and user profiling
PubMed	Biomedical Literature	197,000+	High	Medical and health-related information
AG News	News Classification	127,600	Low–Medium	Topic-sensitive information
CNN/DailyMail	News Summarization	300,000+	Medium	Contextual information leakage
SynthPII	Synthetic Sensitive Records	Custom Generated	Very High	Personally Identifiable Information (PII)

Table 2. Main retrieval performance of No-DP, Static DP (epsilon=5.0), and PADP across datasets

Delta_f	Dataset	No-DP R@1	Static5 R@1	PADP R@1	PADP NDCG@10	PADP eps_eq
0.05	IMDB	0.146	0.130	0.144	0.614	9.43
0.05	PubMed	0.120	0.106	0.118	0.478	10.00
0.05	AG News	0.130	0.130	0.136	0.661	3.51
0.05	CNN/DM	0.096	0.062	0.072	0.370	3.09
0.05	SynthPII	0.156	0.132	0.154	0.676	3.99
0.15	IMDB	0.146	0.078	0.140	0.594	9.43
0.15	PubMed	0.120	0.056	0.114	0.470	10.00
0.15	AG News	0.130	0.042	0.126	0.602	3.51
0.15	CNN/DM	0.096	0.012	0.052	0.274	3.09
0.15	SynthPII	0.156	0.046	0.136	0.644	3.99

R@1=Recall@1; Static5=Static DP with epsilon=5.0; eps_eq=equivalent reported privacy budget. Values are means over seeds and queries.

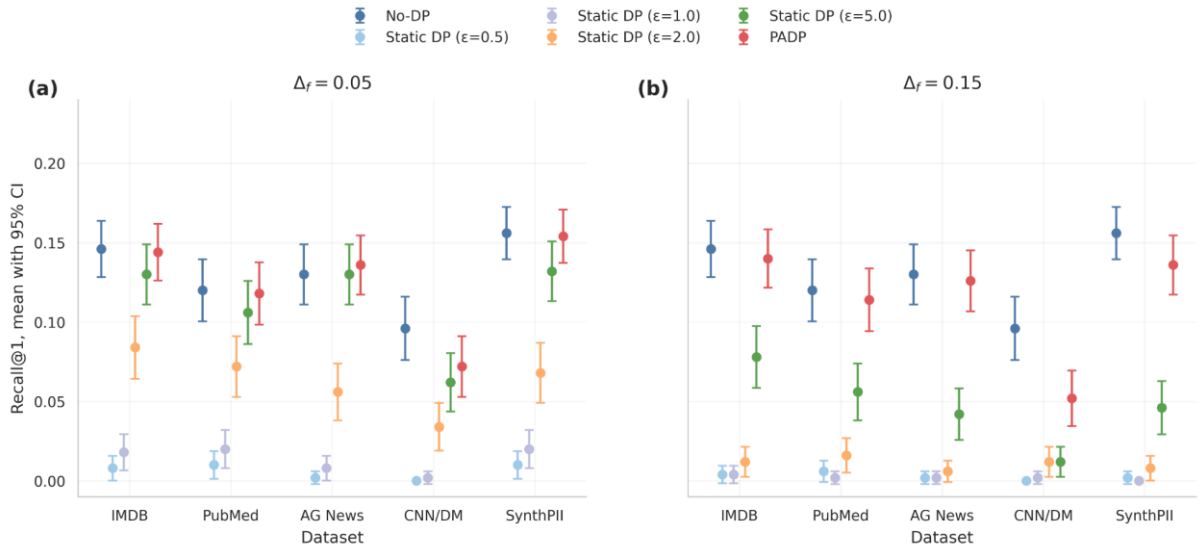

Figure 2. Retrieval performance under two perturbation levels. Points represent mean Recall@1 and error bars denote 95% confidence intervals. PADP maintains retrieval utility close to No-DP and substantially above low-budget static DP baselines.

Table 3. PADP improvement over Static DP (epsilon=5.0) on Recall@1

Delta_f	Dataset	PADP	Static5	Abs. gain	Rel. gain (%)	P	Hedges g
0.05	IMDB	0.144	0.130	0.014	10.8	7.07e-02	0.15
0.05	PubMed	0.118	0.106	0.012	11.3	1.34e-01	0.12
0.05	AG News	0.136	0.130	0.006	4.6	5.13e-01	0.06
0.05	CNN/DM	0.072	0.062	0.010	16.1	3.98e-01	0.11
0.05	SynthPII	0.154	0.132	0.022	16.7	2.18e-02	0.24
0.15	IMDB	0.140	0.078	0.062	79.5	3.46e-07	0.65
0.15	PubMed	0.114	0.056	0.058	103.6	3.42e-06	0.61
0.15	AG News	0.126	0.042	0.084	200.0	5.73e-09	0.93
0.15	CNN/DM	0.052	0.012	0.040	333.3	1.57e-04	0.56
0.15	SynthPII	0.136	0.046	0.090	195.7	1.30e-09	1.00

P values are from paired Wilcoxon signed-rank tests. Static5=Static DP with epsilon=5.0.

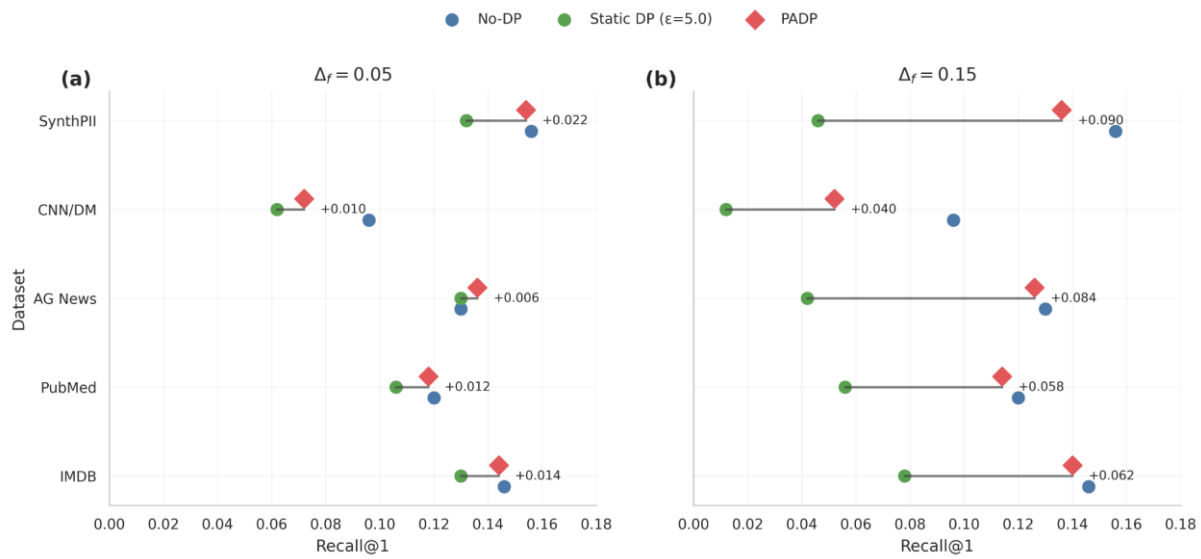

Figure 3. Dataset-level comparison between PADP and Static DP (epsilon=5.0). The dumbbell lines show Recall@1 gains obtained by PADP. Gains are particularly pronounced under Delta_f = 0.15.

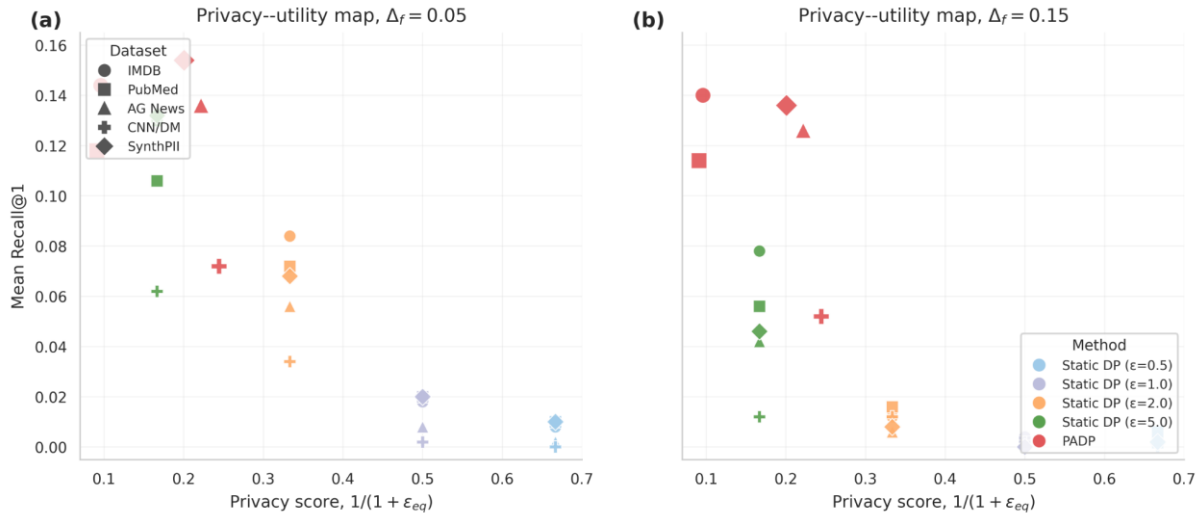

Figure 4. Privacy-utility map. The x-axis shows privacy score 1/(1+epsilon_eq), and the y-axis shows mean Recall@1. PADP provides a favorable operating point compared with static DP baselines.

3.4. Ablation analysis

Table 4 and Figure 5 present ablation outcomes. No adaptive epsilon achieved slightly higher NDCG, whereas Full PADP provided a more balanced trade-off between retrieval effectiveness and adaptive privacy allocation setup. Without adaptive epsilon, the system struggled to prioritize delicate instances more effectively. In contrast, omitting PII weighting dulled the responsiveness of resource distribution based on sensitivity. Average scores dipped consistently under the Laplace version compared to full configuration. Performance relies not on one piece alone - instead, it emerges through coordination of dynamic allocation, sensitivity assessment, and noise injection using Gaussian patterns.

3.5. Membership Distinguishability

Results showing how well membership can be distinguished appear in Table 5 and Figure 6. When KS scores are smaller and AUC values near 0.5, it means detection is harder. In contrast to No-DP, PADP tended to make distributions less detectable, yet outcomes shifted across datasets. On the AG News set, for instance, PADP reached an AUC of 0.410 - nearer to chance than Static DP at epsilon 5.0, where AUC was 0.530. Similar behavior emerged on PubMed and CNN/DM, with PADP maintaining performance around the level expected by randomness. Though SynthPII posed the greatest difficulty, it's PADP AUC exceeding 0.5 reveals persistent issues in masking dense synthetic data without stronger adjustments. Instead of viewing PADP as a universal fix for membership detection, think of it as shifting the trade-off between privacy and usefulness toward better outcomes.

An exception was observed on the SynthPII dataset, where distinguishability remained comparatively higher despite adaptive privacy allocation. This finding suggests that datasets dominated by highly concentrated identifier patterns may require stronger privacy defenses than adaptive perturbation alone.

3.6. Adaptive Epsilon Behavior and Sensitivity Analysis

Despite appearing complex, Figure 7 reveals how PADP adjusts privacy budgets based on need. A downward trend in the scatter plot links higher sensitivity scores to lower epsilon_eq values - indicating tighter privacy for more sensitive cases. Such patterns align closely with the logic laid out in equation 2. Variation across data becomes evident when examining Figure 8 alongside Table 6, where ideal lambda settings shift depending on the dataset used. Looking across datasets, PubMed worked best at lambda = 1. Meanwhile, IMDB showed stronger results with lambda = 2. For AG News, values between 5 and 10 delivered better outcomes. In contrast, CNN/DM performed optimally around lambda = 10. On the synthetic data - SynthPII - lambda ranging from 2 to 5 was preferable. Taken together, these patterns suggest a fixed privacy setting does not suit varied text collections equally well.

Table 4. Ablation summary averaged across datasets

Variant	Recall@1	NDCG@10	Mean eps	Mean sigma
Full PADP	0.118	0.483	28.54	0.034
No adaptive eps	0.114	0.498	48.45	0.015
No PII weighting	0.094	0.416	18.59	0.056
Laplace variant	0.096	0.415	5.89	0.048

Values are averages across datasets. eps= privacy budget before equivalent-scale conversion in ablation output.

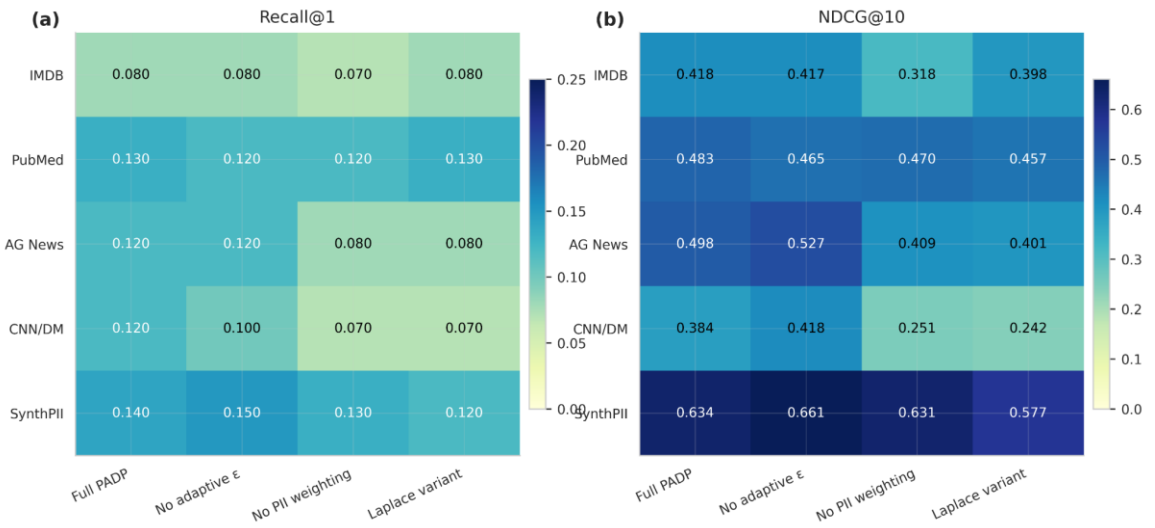

Figure 5. Ablation analysis for Recall@1 and NDCG@10. Full PADP generally provides the most balanced performance, while removing PII weighting or using the Laplace variant reduces ranking utility.

Table 5. Membership distinguishability assessment for No-DP, Static DP (epsilon=5.0), and PADP

Dataset	Method	KS	AUC	|AUC-0.5|	Risk
IMDB	No-DP	0.400	0.609	0.109	Elevated
IMDB	Static DP (eps=5.0)	0.200	0.561	0.061	Elevated
IMDB	PADP	0.300	0.578	0.078	Elevated
PubMed	No-DP	0.200	0.473	0.027	Elevated
PubMed	Static DP (eps=5.0)	0.233	0.407	0.093	Elevated
PubMed	PADP	0.200	0.464	0.036	Elevated
AG News	No-DP	0.433	0.291	0.209	Elevated
AG News	Static DP (eps=5.0)	0.200	0.530	0.030	Elevated
AG News	PADP	0.267	0.410	0.090	Elevated
CNN/DM	No-DP	0.333	0.372	0.128	Elevated
CNN/DM	Static DP (eps=5.0)	0.200	0.457	0.043	Elevated
CNN/DM	PADP	0.200	0.406	0.094	Elevated
SynthPII	No-DP	0.167	0.502	0.002	Elevated
SynthPII	Static DP (eps=5.0)	0.167	0.480	0.020	Elevated
SynthPII	PADP	0.267	0.582	0.082	Elevated

KS= Kolmogorov-Smirnov statistic; AUC closer to 0.5 indicates weaker distinguishability.

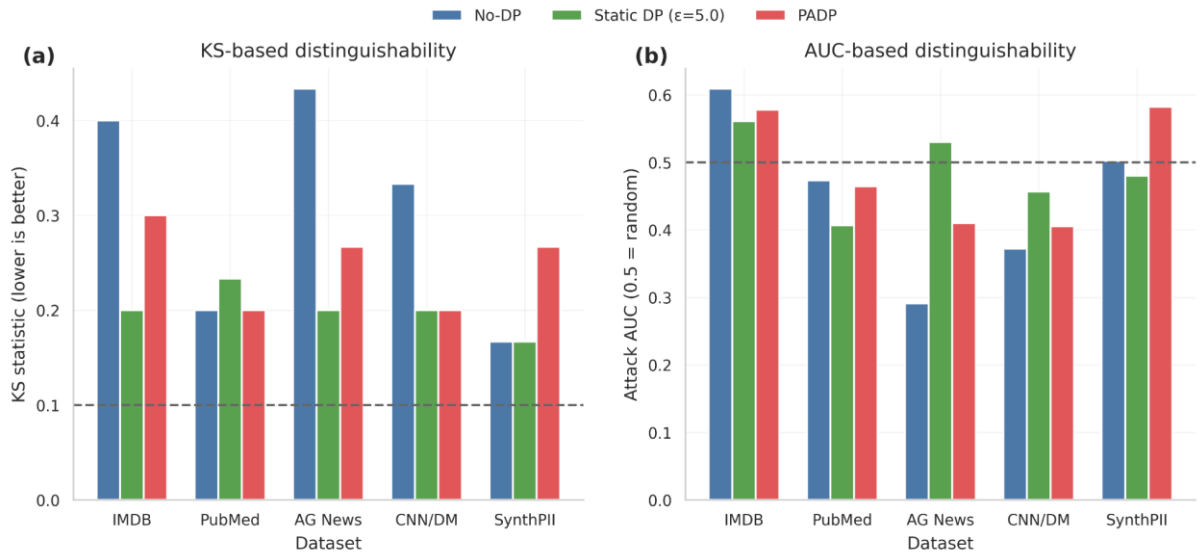

Figure 6. Membership distinguishability assessment using KS statistic and AUC. Lower KS and AUC closer to 0.5 indicate weaker distinguishability. PADP provides competitive privacy-risk behavior while preserving utility.

Table 6. Sensitivity analysis summary across T_th values

Dataset	Optimal lambda range	Mean PU	Mean privacy score	Mean Recall@1
IMDB	2	0.113	0.094	0.140
PubMed	1	0.116	0.091	0.160
AG News	5-10	0.176	0.266	0.133
CNN/DM	10	0.140	0.242	0.100
SynthPII	2-5	0.180	0.251	0.145

PU=privacy-utility score. Privacy score is computed as 1/(1+epsilon_eq).

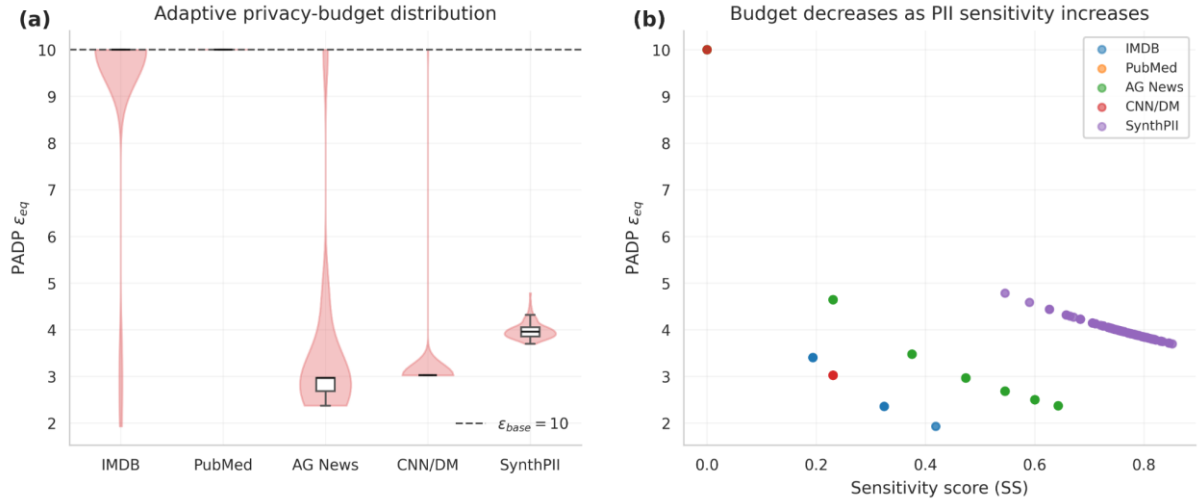

Figure 7. Adaptive privacy-budget behavior of PADP. Higher sensitivity scores lead to lower effective epsilon values, confirming the intended sensitivity-aware allocation mechanism.

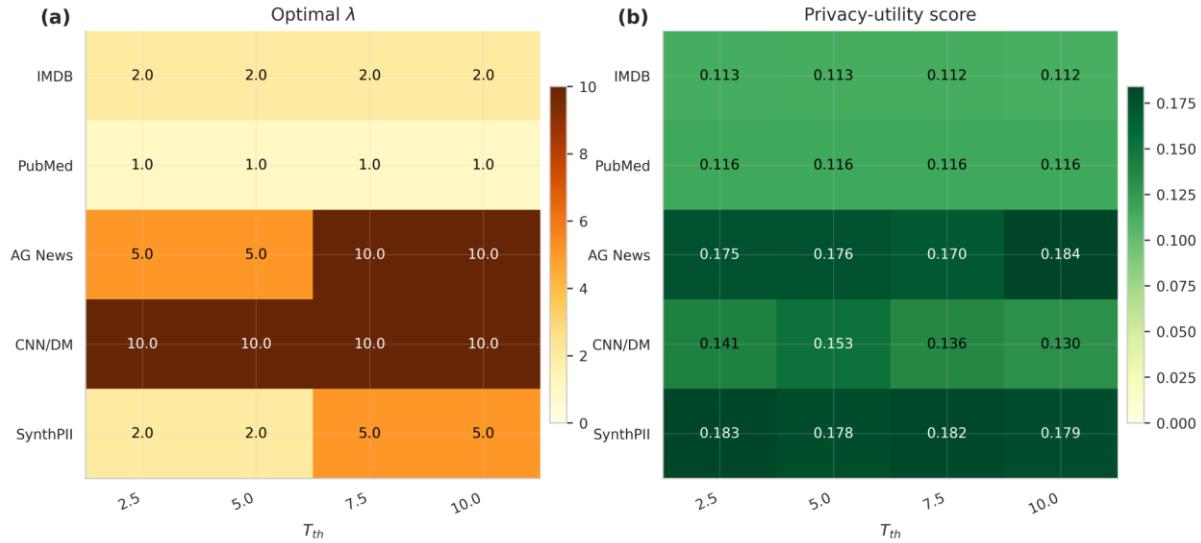

Figure 8. Sensitivity analysis of optimal lambda and privacy-utility score across T_th values. The results show that optimal adaptation strength varies across datasets.

4. Discussion

What stands out is how adaptive privacy distribution cuts down the drop in usefulness compared to standard fixed DP methods. This gain becomes clearest when Delta_f hits 0.15, a point at which constant-budget approaches falter badly. Instead of distorting every input equally, PADP eases the burden on privacy spending - especially targeting those data points more easily exposed. By shifting focus toward sensitivity differences, it avoids one-size-fits-all noise application. Each adjustment responds dynamically, preserving accuracy without sacrificing protection. Notably, performance dips slow where older models struggle most.

Results align with existing work on privacy in machine learning, where balancing strong data protection and functional models remains a known challenge (Dwork et al.,, 2006; Abadi et al., 2016). Unlike conventional differential privacy approaches, PADP adjusts embeddings during retrieval while allocating privacy budgets based on sensitivity. Because of this, it functions better as an adaptable supplement within retrieval frameworks instead of substituting complete differentially private training pipelines.

One notable finding stands out: PADP achieved top performance on AG News and SynthPII. Heterogeneous semantic categories appear in one dataset; the other holds tightly packed synthetic identifiers instead. When content varies like this, fixed perturbations may overwhelm safer material or underprotect sensitive segments. Flexibility becomes critical - adaptive methods adjust intensity based on context. Still, results shift when looking at PubMed, where changes across versions fade. Structural consistency marks these biomedical summaries, plus personal identifier traces run sparse through the examples drawn.

Although membership distinguishability findings need cautious reading, PADP demonstrated favorable privacy–utility behavior under the evaluated settings. While

outperforming Static DP across several measures, it falls short on certain risk indicators. That outcome makes sense - intensifying fixed noise may limit identification yet harms data value too much. Instead, PADP focuses on keeping outputs meaningful without sacrificing strong protection levels. Such a balance fits real-world search platforms where degraded performance undermines adoption.

An important observation is that PADP did not consistently improve distinguishability metrics across all datasets. In particular, SynthPII exhibited increased distinguishability under PADP, suggesting that adaptive perturbation alone may be insufficient in densely sensitive corpora.

The different behavior on the SynthPII dataset can be explained by the structural properties of this corpus. SynthPII, unlike natural language datasets, contains dense, repetitive, and highly concentrated personally identifiable information patterns. In such a case, many documents might get equally high sensitivity scores; hence, the relative advantage of adaptive privacy-budget allocation is reduced. When sensitivity values are located within a narrow high-risk range, PADP does not have much room to distinguish between low-risk and high-risk documents.

This might be an explanation for why PADP gave more membership distinguishability on SynthPII than what was achieved with Static DP. The finding here seems to point out that adaptive perturbation by itself might not be enough in cases where repetitive identifier structures dominate the dataset. Stronger mechanisms for preserving privacy would be necessary in such cases—these could include more aggressive perturbation, formal privacy accounting, synthetic data generation, or hybrid protection strategies.

Furthermore, runtime overhead was not comprehensively evaluated. Although PADP was designed as middleware, operational deployment characteristics require dedicated benchmarking.

One limitation lies in how membership analysis operates - built on similarity-based distinction, it does not reflect resilience toward stronger attacks like LiRA or methods using shadow models (Emelyanov et al., 2026). Instead of measuring broad vulnerability, it captures only narrow separation cues. The sensitivity metric brings another constraint: its outcome ties closely to how well PII is spotted and how weights are assigned. Depending too much on detection accuracy, it may shift if tools improve. Testing stayed within text-based retrieval contexts, leaving out multimodal setups entirely (Zhang et al., 2025). Systems that generate language or learn across decentralized devices were simply not part of any trial. Ending with full privacy tracking across entire model development still falls beyond current reach. From these gaps, clear paths emerge toward what comes next.

The present study does not establish a strict database-level differential privacy guarantee for adaptive retrieval mechanisms.

Overall, PADP should not be interpreted as a universally superior privacy mechanism. Rather, it represents a practical retrieval-time strategy that seeks to improve privacy–utility trade-offs under realistic deployment constraints. Establishing formal guarantees for adaptive retrieval mechanisms and validating them against stronger adversaries remain important directions for future work.

Despite these limitations, the experimental findings consistently demonstrate that adaptive privacy allocation provides a practical and effective mechanism for improving the privacy–utility balance in sensitive retrieval environments. The identified limitations therefore represent opportunities for future refinement rather than fundamental weaknesses of the proposed framework.

4.1. Practical Implications

The PADP framework has practical implications for privacy-aware retrieval systems operating in the wild. Since PADP is a retrieval-time middleware, it can be inserted into existing RAG pipelines without requiring retraining of the underlying large language model or reconstruction of the vector database. Hence, it fits enterprise search systems, healthcare-oriented RAG applications, legal document retrieval platforms, and corporate knowledge bases where sensitive information might share space with general-purpose documents (Martinelli and Ponti, 2026).

In such environments, applying a uniform perturbation strategy to all documents may unnecessarily reduce retrieval utility. PADP addresses this limitation by assigning stronger perturbation to documents with higher PII-based sensitivity while preserving semantic retrieval performance for lower-risk content. Therefore, the framework can support practical privacy–utility management in heterogeneous document repositories. However, because PADP does not provide formal database-level Differential Privacy guarantees, it should be used as a complementary privacy-enhancement layer rather than as a standalone formal privacy mechanism.

5. Conclusion

This work presented PADP, a retrieval-time adaptive perturbation framework intended to improve the balance between retrieval effectiveness and operational privacy enhancement in RAG systems. Instead of fixed rules, it shifts protection levels using document-specific risk indicators along with emphasis on personally identifiable details. Testing occurred across five collections, applying two distinct noise methods, where results showed stronger recall performance at top rank versus unchanging privacy models. Gains stood out most when tighter constraints were applied, specifically with Delta_f set to 0.15.

Among the tested benchmarks, SynthPII led with a gain of 0.090 above Static DP at epsilon 5.0; AG News followed closely with an increase of 0.084. Performance rose by 0.062 on IMDB, whereas PubMed recorded a lift of 0.058

- CNN/DM reached +0.040. When breaking down contributions, removing either adaptive epsilon assignment or sensitivity-based weighting weakened outcomes consistently. Instead of pooling results uniformly, shifts in risk profiles emerged under membership inference scrutiny. Rather than fixed settings, best configurations shifted meaningfully depending on data origin.

Still, PADP offers a straightforward way to better manage privacy and usefulness in searching sensitive texts. Moving ahead, research could include tougher tests against attacks, clearer methods to track privacy loss, while also exploring uses in systems that boost responses using retrieved data or rely on large language models.

Author Contributions

The percentages of the authors' contributions are presented below. All authors reviewed and approved the final version of the manuscript.

	S.M.	Y.V.	Ö.C.T.
C	70	20	10
D	70	20	10
S	20	40	40
DCP	60	20	20
DAI	60	20	20
L	60	30	10
W	70	20	10
CR	40	30	30
SR	40	30	30
PM	40	30	30
FA	40	30	30

C= concept, D= design, S= supervision, DCP= data collection and/or processing, DAI= data analysis and/or interpretation, L= literature search, W= writing, CR= critical review, SR= submission and revision, PM= project management, FA= funding acquisition.

Conflict of Interest

The authors declared that there is no conflict of interest.

Ethical Consideration

Ethics committee approval was not required for this study because of there was no study on animals or humans.

Acknowledgements

This article was written as part of a doctoral dissertation at the Istanbul University-Cerrahpasa Institute of Graduate Studies.

References

Anderson, M., Amit, G., & Goldsteen, A. (2025). Is my data in your retrieval database? Membership inference attacks against retrieval augmented generation. In *Proceedings of the International Conference on Information Systems Security and Privacy* (pp. 345–354). SCITEPRESS. https://doi.org/10.5220/0013108300003899

Azzam, R., Musamih, A., Gebreab, S. A., Salah, K. H., & Omar, M. A. (2026). Agentic LLM for anonymizing healthcare data with contextual awareness. *Knowledge-Based Systems*, *312*, Article 116034. https://doi.org/10.1016/j.knosys.2026.116034

Castagnaro, A., Salviati, U., Conti, M., Pajola, L., & Pizzi, S. (2025). The hidden threat in plain text: Attacking RAG data loaders. In *Proceedings of the 18th ACM Workshop on Artificial Intelligence and Security* (AISec 2025) (pp. 45–56). ACM. https://doi.org/10.1145/3733799.3762976

Chen, B., Han, N., & Miyao, Y. (2025). A statistical and multi-perspective revisiting of the membership inference attack in large language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (pp. 1124–1138). ACL. https://doi.org/10.18653/v1/2025.acl-long.1114

Dwork, C., McSherry, F., Nissim, K., & Smith, A. D. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference* (pp. 265–284). Springer. https://doi.org/10.1007/11681878_14

Emelyanov, A. A., Kudriashov, S., & Alena, F. S. (2026). FiMMIA: Scaling semantic perturbation-based membership inference across modalities. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics* (EACL 2026) (pp. 89–94). ACL. https://doi.org/10.18653/v1/2026.eacl-demo.11

Eyupoglu, C., Aydin, M. A., Zaim, A. H., & Sertbas, A. (2018). An efficient big data anonymization algorithm based on chaos and perturbation techniques. *Entropy*, *20*(5), Article 373. https://doi.org/10.3390/e20050373

Ferrag, M. A., Tihanyi, N., Hamouda, D., Lakas, A., & Debbah, M. A. (2026). From prompt injections to protocol exploits: Threats in LLM-powered AI agents workflows. *ICT Express*, *12*(1), 15–28. https://doi.org/10.1016/j.icte.2025.12.001

Gade, S. R., Ponna, D., Shah, J., & Patel, K. A. (2026). Retrieval-Augmented Generation (RAG): Architecture, ethical challenges, and optimization strategies. In *Proceedings of the 3rd International Conference on Machine Learning and Autonomous Systems* (ICMLAS 2026) (pp. 204–211). IEEE. https://doi.org/10.1109/ICMLAS67792.2026.11483907

Gao, X., Meng, X., Dong, Y., Li, Z., & Guo, S. (2025). DCMI: A differential calibration membership inference attack against Retrieval-Augmented Generation. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security* (CCS 2025) (pp. 1421–1435). ACM. https://doi.org/10.1145/3719027.3765103

Ge, Y., Zhang, H., Sun, H., Huang, Z., & Yang, H. (2026). PrivTabRAG: A data-model co-design framework for privacy-preserving retrieval-augmented generation on tabular data. *Expert Systems with Applications*, *264*, Article 132023. https://doi.org/10.1016/j.eswa.2026.132023

Geng, T., Xu, Z., Qu, Y., & Wong, W. E. (2026). Prompt injection attacks on large language models: A survey of attack methods, root causes, and defense strategies. *Computers, Materials & Continua*, *82*(3). https://doi.org/10.32604/cmc.2025.074081

Grislain, N. (2025). RAG with differential privacy. In *Proceedings of the 2025 IEEE Conference on Artificial Intelligence* (CAI 2025) (pp. 112–119). IEEE. https://doi.org/10.1109/CAI64502.2025.00150

Habernal, I. (2021). When differential privacy meets NLP: The

devil is in the detail. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP 2021) (pp. 1522–1528). ACL. https://doi.org/10.18653/v1/2021.emnlp-main.114

Habernal, I. (2022). How reparametrization trick broke differentially-private text representation learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (pp. 671–676). ACL. https://doi.org/10.18653/v1/2022.acl-short.87

He, L., Tang, P., Zhang, Y., Zhou, P., & Su, S. (2025). Mitigating privacy risks in Retrieval-Augmented Generation via locally private entity perturbation. *Information Processing & Management*, *62*(3), Article 104150. https://doi.org/10.1016/j.ipm.2025.104150

He, Y., Li, B., Liu, L., Ren, K., & Chen, C. (2025). Towards label-only membership inference attack against pre-trained large language models. In *Proceedings of the 34th USENIX Security Symposium*. USENIX Association.

Li, J., Wang, J., Zhao, Z., & Wei, Z. (2026). Large language models in practice: A comprehensive review of privacy risks, attacks, and defense mechanisms for intelligent system deployment. *Computer Standards & Interfaces*, *94*, Article 104181. https://doi.org/10.1016/j.csi.2026.104181

Li, Y., Liu, G., Wang, C., & Yang, Y. (2025). Generating is believing: Membership inference attacks against Retrieval-Augmented Generation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP 2025) (pp. 4120–4124). IEEE. https://doi.org/10.1109/ICASSP49660.2025.10889013

Mainetti, L., & Elia, A. (2025). Detecting personally identifiable information through natural language processing: A step forward. *Applied System Innovation*, *8*(2), Article 55. https://doi.org/10.3390/asi8020055

Maini, P., Jia, H., Papernot, N., & Dziedzic, A. (2024). LLM dataset inference: Did you train on my dataset? In *Advances in Neural Information Processing Systems* (NeurIPS 2024). Curran Associates.

Martinelli, I., & Ponti, M. A. (2026). Towards secure Retrieval-Augmented Generation: Preventing LLM data leaks with RBAC. In *International Conference on Security and Privacy in Communication Networks* (pp. 512–527). Springer. https://doi.org/10.1007/978-3-032-15993-9_34

Moia, V. H. G., Sanz, I. J., Rebello, G. A. F., Hitaj, B., & Lindqvist, U. (2026). LLM in the middle: A systematic review of threats and mitigations to real-world LLM-based systems. *Computer Science Review*, *59*, Article 100916. https://doi.org/10.1016/j.cosrev.2026.100916

Pan, K., Ong, Y. S., Gong, M., Qin, A. K., & Gao, Y. (2024). Differential privacy in deep learning: A literature survey. *Neurocomputing*, *580*, 234–249. https://doi.org/10.1016/j.neucom.2024.127663

Puerto, H., Gubri, M., Yun, S., & Oh, S. J. (2025). Scaling up membership inference: When and how attacks succeed on large language models. In *Findings of the Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics* (Findings of NAACL 2025) (pp. 312–326). ACL. https://doi.org/10.18653/v1/2025.findings-naacl.234

Qi, Z., Zhang, H., Xing, E. P., Kakade, S. M., & Lakkaraju, H. (2025). Follow my instruction and spill the beans: Scalable data extraction from Retrieval-Augmented Generation systems. In *Proceedings of the 13th International Conference on Learning Representations* (ICLR 2025).

Sanchez-Serrano, P., Rios, R., & Agudo, I. (2025). A decision framework for privacy-preserving synthetic data generation. *Computers and Electrical Engineering*, *122*, Article 110468. https://doi.org/10.1016/j.compeleceng.2025.110468

Szawerna, M. I., Dobnik, S., Sánchez, R. M., Tiedemann, T. L., & Volodina, E. (2024). Detecting personal identifiable information in Swedish learner essays. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization* (CALD-pseudo 2024) (pp. 45–53). ACL.

Vonderhaar, L., MacHado, D. A., & Ochoa, O. (2025). Surveying the RAG attack surface and defenses: Protecting sensitive company data. In *Proceedings of the 2025 IEEE International Conference on Artificial Intelligence Testing* (AITest 2025) (pp. 24–31). IEEE. https://doi.org/10.1109/AITest66680.2025.00015

Vural, Y., & Aydos, M. (2017). A new approach to utility-based privacy preserving in data publishing. In *Proceedings of the 2017 IEEE International Conference on Computer and Information Technology* (CIT) (pp. 204–209). IEEE. https://doi.org/10.1109/CIT.2017.27

Wang, G., He, J., Li, H., Zhang, M., & Feng, D. (2025). RAG-leaks: Difficulty-calibrated membership inference attacks on retrieval-augmented generation. *Science China Information Sciences*, *68*(4), Article 144201. https://doi.org/10.1007/s11432-024-4441-4

Wang, T., Zhai, L., Yang, T., Luo, Z., & Liu, S. (2024). Selective privacy-preserving framework for large language models fine-tuning. *Information Sciences*, *665*, Article 121000. https://doi.org/10.1016/j.ins.2024.121000

Ward, C. M., & Harguess, J. D. (2025). Adversarial threat vectors and risk mitigation for Retrieval-Augmented Generation systems. In *Proceedings of SPIE - The International Society for Optical Engineering*, *12450*, Article 12.3055931. https://doi.org/10.1117/12.3055931

Wu, Y. (2025). Dynamic privacy budget allocation mechanism based on multi-dimensional query management. In *Proceedings of the International Conference on Electrical, Computer, and Energy Technologies* (ICECET 2025). IEEE. https://doi.org/10.1109/ICECET63943.2025.11472026

Wu, Z., Luo, S. L., & Pan, L. (2026). A semantics-maintained differential privacy protection for high-utility text. *Knowledge-Based Systems*, *310*, Article 115478. https://doi.org/10.1016/j.knosys.2026.115478

Xie, R., Wang, J., Huang, R., Gong, N. Z., & Dhingra, B. (2024). RECALL: Membership inference via relative conditional log-likelihoods. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP 2024) (pp. 5621–5635). ACL. https://doi.org/10.18653/v1/2024.emnlp-main.493

Xin, Y., Li, Z., Yu, N., Backes, M., & Zhang, Y. (2024). Inside the black box: Detecting data leakage in pre-trained language encoders. In *Proceedings of the European Conference on Artificial Intelligence* (ECAI 2024). IOS Press. https://doi.org/10.3233/FAIA240960

Yan, B., Li, K., Xu, M., Ren, Z., & Cheng, X. (2025). On protecting the data privacy of Large Language Models (LLMs) and LLM agents: A literature review. *High-Confidence Computing*, *5*(1), Article 100300. https://doi.org/10.1016/j.hcc.2025.100300

Yao, Y., Duan, J., Xu, K., Sun, Z., & Zhang, Y. (2024). A survey on Large Language Model (LLM) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, *4*(2), Article 100211. https://doi.org/10.1016/j.hcc.2024.100211

Ye, H., Guo, J., Liu, Z., & Lam, K. Y. (2025). Efficient privacy-preserving Retrieval Augmented Generation with distance-preserving encryption. In *Proceedings of the 2025 3rd International Conference on Foundation and Large Language Models* (FLLM 2025) (pp. 74–81). IEEE.

https://doi.org/10.1109/FLLM67465.2025.11391120

Zeng, H., Liu, X., Hu, Y., Tang, S., & Chen, G. (2026). Automated annotation of privacy information in user interactions with large language models. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1420–1434). ACM. https://doi.org/10.1145/3770854.3785680

Zhang, J., Zeng, S., Ren, J., Tang, X., & Chang, Y. (2025). Beyond text: Unveiling privacy vulnerabilities in multi-modal Retrieval-Augmented Generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP 2025) (pp. 1245–1259). ACL. https://doi.org/10.18653/v1/2025.emnlp-main.1259