A Survey on Privacy-Preserving Multimodal AI Assistant Using Local Retrieval-Augmented Generation and Selective Cloud Processing

Wajid Hussain
Department of Computer Science, Lahore Leads University, Lahore Pakistan
*Correspondence: wajid.hussain0335@gmail.com

Citation | Hussain. W, "A Survey on Privacy-Preserving Multimodal AI Assistant Using Local Retrieval-Augmented Generation and Selective Cloud Processing", IJIST, Vol. 8 Issue. 9 pp 3958-3976, October 2026

Received | Oct 12, 2026 **Revised |** Oct 27, 2025 **Accepted |** Oct 29, 2026 **Published |** Oct 31, 2026.

The fast development of multimodal Artificial Intelligence (AI) assistants allow users to interact with text, images, documents, and audio among a single intelligent system. However, existing AI assistants often rely heavily on cloud-based processing, requiring sensitive user data to be transmitted to remote servers. This prompts important concerns regarding privacy leakage, data security, communication overhead, and dependence on external cloud infrastructure. The proposed study aims to address these challenges by developing a Privacy-Preserving Multimodal AI Assistant using Local Retrieval-Augmented Generation (RAG) and Selective Cloud Processing**.** The motivation of this research is to attain a balance between privacy, response quality, computational efficiency, and latency. The proposed methodology will first process multimodal inputs locally utilizing Optical Character Recognition (OCR), speech recognition, document processing, and vision models. Relative information will then be cataloged in a local vector database and retrieved among a local RAG architecture. A privacy assessor will examine sensitive information such as names, identification numbers, email addresses, and medical information. An adaptive privacy-aware router will subsequently assess whether a query should be processed entirely locally or whether sanitized facts can be selectively sent to a cloud-based AI model. A final privacy-verification step will assure that sensitive facts is not unconsciously exposed. The proposed system is an adaptive local-cloud multimodal RAG architecture that dynamically equalizes privacy and AI performance. This technique will be evaluated using answer accuracy, RAM/VRAM, retrieval performance, computational resources, and privacy leakage rate (PLR), latency, and cloud data transmission. The importance of this research lies in providing a practical design for deploying trustworthy multimodal AI assistants in privacy-sensitive environments such as healthcare, education, research, and enterprise applications.

Keywords: Multimodal Artificial Intelligence; Local Retrieval-Augmented Generation; Optical Character Recognition; Selective Cloud Processing and Privacy-Preserving

Introduction:

The rapid development of large language models (LLMs) and multimodal artificial intelligence (MAI) has significantly amplified the capabilities of intelligent systems for understanding, retrieving and processing information from heterogeneous data sources. Recent AI assistants are rapidly significant to process not only textual queries but also scanned documents, tables, images, PDFs, and audio. This development has produced opportunities for intelligent applications in healthcare, education, scientific research, finance, business, and personal productivity. However, the increasing use of heterogeneous and potentially sensitive data also addresses important challenges related to privacy, data governance, computational requirements, and fence deployment. Modern study on trustworthy RAG focuses that explainability, reliability, safety, privacy, and accountability must be considered together when deploying RAG-based AI systems in practical environments [1][2]. RAG has emerged as an important approach for improving the factual grounding and knowledge-access capabilities of generative AI. Other than relying exclusively on knowledge encoded in model parameters, RAG fetches related information from an external knowledge source and incorporates the fetched information into the generation process. This framework permits an AI system to access domain-specific or dynamically updated information without involving complete retraining of the underlying language model. Modern research has further explored privacy-aware RAG architectures and demonstrated that privacy protection must be considered throughout the retrieval, context construction, and generation stages rather than only at the final output stage [2][3]. In particular, recent privacy-preserving RAG research has explored fine-grained privacy processing and context-aware evaluation to improve the balance between information utility and privacy protection [3].

The development of multimodal AI has extended the conventional text-based RAG paradigm toward Multimodal Retrieval-Augmented Generation (MRAG)**.** Multimodal RAG permits AI systems to fetch and reason over heterogeneous information such as text, images, documents, and audio. Surveys of multimodal RAG have detected cross-modal retrieval, representation alignment, multimodal fusion, reasoning, and evaluation as an important challenges in this emerging research area [4]. Previously work such as MuRAG justified the value of combining image and textual information within a retrieval-augmented design, representing that external multimodal memory can enhance question answering over multimodal knowledge [5]. Subsequent research has expanded multimodal retrieval toward document understanding and other modalities, justifying that visual and structural information contained in documents may be lost when documents are converted exclusively into plain text [6]. The development from text-only RAG toward multimodal RAG, though, also amplifies the privacy attack surface. Multimodal systems may process medical information, identification numbers, confidential documents, names, addresses, voice recordings, photographs, and other personally identifiable information (PII). In consequence, privacy risks are no longer restricted to textual retrieval databases. Novel research has validated that multimodal RAG systems can flatten sensitive information throughout both direct reproduction and indirect descriptions of retrieved multimodal content. Especially, attacks against vision-language and speech-language RAG systems verify that multimodal information can suggest privacy risks that necessitate reliable protection approaches [7]. These rulings signify that privacy preservation should be incorporated into the architecture before sensitive multimodal information is transmitted to external processing environments.

Privacy outlets are also evident in conventional RAG systems. Empirical researches have authenticated that retrieved private information can be revealed through carefully designed queries, exposing that RAG may establish privacy risks in addition to the risks already associated with LLMs [8]. These findings are particularly important for applications in which the retrieval database contains proprietary, medical, financial, educational, or personally

identifiable information. Privacy-preserving approaches have therefore investigated techniques such as synthetic data generation, differential privacy, entity perturbation, filtering, and privacy-aware retrieval. For example, SAGE examined the use of synthetic retrieval data to minimize privacy threats while preserving useful contextual information for RAG generation [9]. Such techniques verify the possibility of protecting sensitive retrieval information, but they may require additional preprocessing or may alter the original information before it influences the generation model. Another important direction is privacy-preserving cloud-based RAG. Cloud services provide access to powerful language and multimodal models without requiring users to maintain expensive local computational infrastructure. However, sending queries, retrieved contexts, embeddings, or multimodal data to a cloud provider can create additional privacy and communication concerns. Remote RAG, for example, explicitly investigates privacy-preserving cloud RAG and addresses the risk associated with transmitting user queries to remote retrieval services [10]. Similarly, federated RAG research has examined the combination of federated learning and RAG to reduce direct exposure of distributed data while supporting knowledge-intensive applications. A recent systematic mapping study identified privacy-preserving retrieval, client heterogeneity, and evaluation as important challenges in federated RAG [11]. Nevertheless, federated RAG primarily addresses distributed learning and retrieval scenarios and does not directly provide an inference-time mechanism that dynamically decides whether an individual multimodal request should remain local or be processed by a sanitized cloud service. Recent research has also highlighted the importance of adaptive and context-dependent privacy protection. Static privacy policies may treat all documents or retrieved information equally even though the actual privacy risk can depend on the user's query and the particular information being accessed. This research proposed a prompt-aware dynamic privacy technique in which privacy risk is estimated according to both the sensitivity of retrieved information and it's related to the current query [12]. This direction is specifically related to the proposed study because it supports the idea of dynamically adjusting the level of privacy protection rather than applying the same protection scheme to every request. Similarly, modern privacy-preserving RAG research has presented fine-grained privacy processing at entity and relation levels and has proposed loyal metrics for determining information extraction and personal identification risks [3].

Multimodal document processing contributes an additional objection for privacy-aware AI assistants. Real-world documents consistently contain scanned text, tables, figures, signatures, visual layouts, and other information that may not be sufficiently described by conventional text extraction. VisRAG addressed this restriction by utilizing vision-language representations to fetch multimodal documents while preserving visual and layout information [6]. Multimodal RAG inspections likewise determine heterogeneous modalities and cross-modal reasoning as significant areas for future development [4]. Therefore, a practical privacy-preserving multimodal assistant should serve modality-specific preprocessing locally, including visual processing for images, OCR for scanned documents, document parsing and speech recognition for audio, before examining whether any information requires to be transmitted externally. The central research objection is therefore a privacy-utility-efficiency trade-off. A completely cloud-based framework can assign access to powerful models and substantial computational resources, but it may need transmitting sensitive user fact to external services. Conversely, a completely local architecture can minimize external data transmission and provide stronger control over private information, but multimodal models, vector databases, OCR systems, speech recognition models, and local LLMs may impose considerable computational and memory requirements. Current research work on privacy and security in RAG determines centralized, on-device, federated, and hybrid deployment architectures as distinct techniques with various privacy, utility, and resource characteristics [2]. Consequently,

neither fully local nor fully cloud-based processing necessarily feeds an optimal solution for all types of multimodal queries. Based on these observations, there is a research juncture to explore an adaptive privacy-aware local-cloud framework for multimodal RAG. The proposed research addresses this juncture through a privacy-preserving multimodal AI assistant that prosecutes multimodal preprocessing and local retrieval, detects sensitive information, assesses privacy risk, and dynamically prosecutes whether a request should be processed locally or forwarded to the cloud after sanitization. Low-risk information can predict within the local environment, whereas information necessitating external model competencies can be sanitized over masking, removal, or alternate before selective cloud processing. A final privacy-verification step can then analyze whether sensitive information awaits in the transmitted context or generated response. The proposed framework accordingly assimilates five complementary capabilities: multimodal local processing, local RAG, privacy detection and sanitization, adaptive local-cloud routing, and quantitative privacy-utility evaluation. This assimilated framework contrasts from techniques that point exclusively on multimodal retrieval, privacy attacks, synthetic data, federated learning, or cloud privacy. Existing literature provides important individual components of the problem, but the combination of heterogeneous multimodal sanitized cloud processing, sensitivity-aware routing, inputs, local retrieval, and latency, accuracy, retrieval quality, systematic evaluation of privacy, and communication cost remains insufficiently investigated [3][7][10][11].

The selected problem is consequently significant because future AI assistants are likely to regulate throughout heterogeneous personal and organizational data while simultaneously needing high-quality generative capabilities. The proposed study sights to examine whether selective cloud processing can provide additional model abilities without unnecessarily exploding sensitive information. The experimental research will compare cloud-only RAG, local-only RAG, conventional hybrid RAG, and the proposed adaptive local-cloud multimodal RAG utilizing public multimodal datasets and handled privacy scenarios. Assessment will consider latency, retrieval performance, sensitive-data transmission, answer quality, communication overhead, privacy leakage, and computational resource utilization. Finally, the research pursues to introduce a practical framework for privacy-preserving multimodal AI assistants that dynamically balance privacy protection with model use and computational efficiency. Rather than scheming that all information should be processed locally or that all requests should be sent to the cloud, the proposed approach identifies whether privacy-aware adaptive routing can provide a more adjustable deployment technique. Such an architecture may support future applications in privacy-sensitive environments, including scientific research, education, enterprise information management, healthcare, and other domains in which users require both intelligent multimodal assistance and greater control over their data.

Research Problem:

The research problem is the absence of an practically specified, endeavored multimodal RAG workflow that mutually (i) processes heterogeneous inputs locally, (ii) detects direct and indirect sensitive information, (iii) distributes transparent sensitivity and risk scores, (iv) prefers local or sanitized-cloud inference utilizing explicit thresholds, and (v) analyzes latency, retrieval, communication, privacy, answer quality, and compute cost under a common comparison protocol. Existing researches feed advances in RAG synthetic data, privacy, federated retrieval, multimodal retrieval, and adaptive edge-cloud inference, but these strands are often analyzed with various assumptions, datasets, and metrics [13][8]. PALC-MRAG manipulates this gap throughout a unified experimental protocol and four-way baseline comparison.

Research Questions:

This research meets the following questions:

RQ1: How appropriately can a local multimodal RAG system process and retrieve information from images, PDFs, audio, text, and scanned documents while maintaining user privacy?
RQ2: How exactly can a privacy evaluator detect and determine sensitive information in multimodal user queries and retrieved theme?
RQ3: How appropriately can an adaptive routing technique identify whether a query should be processed locally or selectively throughout a sanitized cloud service?
RQ4: How does the proposed privacy-preserving local-cloud multimodal RAG architecture relate with cloud-only, local-only, and conventional hybrid techniques in terms of retrieval performance, accuracy, privacy leakage, and latency?
RQ5: To what limitation can selective cloud processing minimize sensitive-data transmission while preserving acceptable AI response quality and computational efficiency?

Research Objectives:

The objectives are organized with the research questions and experimental plan:
RO1. Implement a local multimodal ingestion and retrieval pipeline for scientific PDFs, scanned documents, images, text, and audio utilizing the selected public datasets.
RO2. Determine PII detection utilizing F1, recall, and precision, solely reporting findings by entity type and modality.
RO3. Implement a sensitivity-aware routing mechanism with documented risk thresholds and determine its false-cloud decisions, false-local decisions, and routing decisions.
RO4. Relate four system configurations employing retrieval, resource metrics, privacy, latency, answer-quality, and communication under identical test queries.
RO5. Measure privacy-utility trade-offs throughout threshold sensitivity analysis and ablation experiments.
RO6. Release a reproducibility record containing software versions, configuration files, dataset versions, split rules, model identifiers, and statistical procedures, subject to dataset licenses and API terms.

Research Novelty and Contributions:

Table 1. Summarizes the Conceptual Distinction

Approach / representative work	Main focus	Limitation relative to this study	PALC-MRAG distinction
Multimodal RAG: MuRAG; multimodal RAG surveys	Cross-modal retrieval and answer generation	Privacy-aware routing is not the central evaluation target	Adds sensitivity scoring, sanitization, and transmission auditing
RAG privacy and attack studies	Leakage analysis, privacy risks, defenses	May focus on specific attacks or privacy mechanisms rather than the full routing pipeline	Evaluates privacy across ingestion, retrieval, transmission, and output
Synthetic-data / remote / federated RAG	Reduce exposure or distribute retrieval	Different threat models and deployment assumptions; not necessarily request-level local/cloud routing	Compares routing against fixed deployment baselines under a common protocol
Adaptive edge–cloud inference	Allocate inference across compute locations	May not include multimodal RAG retrieval and PII checks in the same pipeline	Jointly evaluates retrieval quality, privacy leakage, latency, and data volume

PALC-MRAG is a proposed integration and evaluation architecture; its novelty does not lie in the invention of multimodal retrieval, RAG, edge-cloud routing or PII detection

individually. Its voluntary contribution is to connect these components in a single definable pipeline and determine the resulting privacy-utility trade-off under consistent conditions.
Unified modality-aware pipeline: PDF/document, text, audio, and image inputs are normalized locally and described with modality-appropriate encoders before retrieval.
Explicit privacy gate: risk assessment is applied to user queries and retrieved context before any external transmission, followed by verification of the transmitted payload and response.
Adaptive routing: the destination is selected using privacy risk, local resource availability, and whether cloud use is permitted-not only a fixed local/cloud configuration.
Four-way comparative evaluation: local-only, cloud-only, conventional hybrid, and proposed adaptive hybrid are tested with shared query sets and metrics.
Reproducible privacy scenarios: controlled synthetic identifiers and seeded transformations allow privacy tests without collecting real personal data.

It is a qualitative synthesis of the cited literature, not a claim that every implementation of each family has identical behavior.

Related Work:

RAG has developed into a significant framework for enhancing the knowledge coverage, factuality, and adaptability of LLMs by combining information retrieval with generative modeling. The literature has gradually moved from conventional text-based RAG toward federated RAG, privacy-preserving RAG, multimodal retrieval, and adaptive cloud-edge inference. This progression is especially related to the proposed research, "Privacy-Preserving Multimodal AI Assistant Using Local Retrieval-Augmented Generation and Selective Cloud Processing," because the suggested system combines these study directions into a unified local-cloud architecture.

Advanced research has rapidly focused on the privacy, security, and deployment challenges of RAG. The study [2] represented a comprehensive survey of security and privacy in RAG and reclassified RAG deployment into centralized, on-device, federated, and hybrid architectures. Their research examines several risk surfaces, including retrieval indices, query information, context construction, generation, and distributed updates, and discusses attacks such as membership inference, index inference, poisoning, gradient leakage, and collusion. Importantly, the survey identifies the privacy-utility trade-off associated with various RAG deployment architectures. This exploration feeds a theoretical foundation for the proposed research because the proposed architecture similarly considers local, cloud, and hybrid processing; however, the proposed work amplifies this perspective by applying adaptive routing specifically to multimodal inputs and by assimilating privacy detection and sanitization before cloud processing.

This research [14] proposed Privacy-Preserving Reasoning with Knowledge-Distilled Parametric RAG (PRAG) to minimized the require to upload plaintext documents to cloud systems. Their approach encodes document knowledge into LoRA parameters, thereby minimizing direct vulnerability of the original document content. The authors determine important confines related to document-specific fine-tuning latency and generalization to out-of-distribution queries. This technique identifies that privacy can be addressed by changing how knowledge is identified rather than simply protecting the communication channel. Though, PRAG does not incorporate the adaptive local-versus-cloud routing technique proposed in this research and is primarily involved with textual document knowledge rather than a unified combination of text, PDFs, images, scanned documents, and audio.

The study [15] explored privacy leakage in multimodal RAG throughout membership-inference and image-caption retrieval attacks. Their experiments explore whether an image is contained in a multimodal RAG database and whether connected metadata can subsequently be extracted. The explorations justify that multimodal RAG systems can expose private information and therefore require loyal privacy-preserving techniques. This work is

particularly significant for the proposed research because it validates that privacy threats expand beyond textual RAG to visual and multimodal retrieval. Though, the study primarily assesses privacy attacks, whereas the proposed research concentrates on privacy prevention throughout local processing, PII detection, sanitization, adaptive routing, and final privacy verification.

This paper [16] introduced PRISM, a privacy-aware cloud-edge LLM inference framework that dynamically determines whether computation should occur at the edge, in the cloud, or collaboratively. PRISM profiles entity-level sensitivity and uses a soft gating mechanism to select an execution mode, followed by adaptive privacy protection for collaborative processing. This research reports enhanced privacy-utility trade-offs together with minimizations in latency and energy consumption compared with uniform privacy approaches. PRISM is highly related to the proposed research because it demonstrates the value of sensitivity-aware adaptive routing. However, PRISM concentrates on general LLM inference and cloud-edge association rather than a multimodal RAG pipeline. The proposed research extends this idea by adding local multimodal preprocessing, local vector retrieval, multimodal RAG, sanitization, and multimodal privacy evaluation.

This research work [5] presented Ask in Any Modality, a comprehensive survey of multimodal RAG. This research surveys multimodal RAG datasets, fusion strategies, generation approaches, retrieval methods, benchmarks, evaluation metrics, training techniques, and open challenges. It identifies that multimodal RAG expands conventional RAG by permitting systems to fetch and reason over audio, video, text, images, and other heterogeneous information. The survey justifies the methodological foundation for the multimodal component of the proposed study. Nevertheless, privacy-aware deployment and adaptive local-cloud processing are not the primary focus of the research. The proposed research therefore expands the multimodal RAG literature by presenting privacy detection and sensitivity-based routing into the multimodal retrieval pipeline.

Research work [6] conducted a systematic privacy analysis of multimodal RAG in Beyond Text: Unveiling Privacy Vulnerabilities in Multi-modal Retrieval-Augmented Generation. The authors investigate both vision-language and speech-language RAG and introduce a black-box compositional prompt attack for extracting private information. Their results demonstrate that multimodal models can show sensitive information directly by reproducing retrieved content or indirectly throughout descriptions of retrieved information. This research feeds strong testimony that privacy protection must be though-out through multiple modalities rather than only within text. The proposed research varies by shifting the focus from privacy exposure analysis to privacy-aware system design, utilizing local processing, PII detection, sanitization, adaptive routing, and post-generation privacy verification.

In this paper [7] recommended RemoteRAG, a privacy-preserving cloud RAG service designed to secure user queries while preserving the advantages of remote retrieval. Their technique uses a distance-based differential privacy method to perturb query representations and restrict the cloud search space. The method considers three significant objectives: privacy, retrieval accuracy, and communication efficiency. Experimental findings identify that privacy protection can be attained without significant retrieval degradation while minimize communication overhead. RemoteRAG is related to the proposed selective-cloud component, but its primary concentration is privacy-preserving remote retrieval of textual information. In contrast, the proposed research implements multimodal processing and local retrieval first and sends information to the cloud only when the adaptive routing mechanism identify that cloud processing is appropriate.

This research [8] conducted a systematic mapping research of Federated RAG, coating research published from 2020 to 2025. Federated RAG combines federated learning with retrieval-augmented generation to assist knowledge-intensive applications without directly

revealing raw client data. The study demonstrates privacy-preserving retrieval, client heterogeneity, and evaluation limitations as important research objectives. Federated RAG therefore provides a significant alternative to centralized cloud RAG for privacy-sensitive applications. Though, federated learning usually addresses distributed model training and knowledge sharing rather than the inference-time decision of whether an individual multimodal request should remain local or be selectively sent to a sanitized cloud model. The proposed research therefore complements rather than replaces federated RAG by investigating adaptive local-cloud inference.

Research work [11] developed VisRAG, a vision-language RAG architecture for multimodal documents. Instead of converting documents perfectly into text before retrieval, VisRAG directly defines document pages as visual information utilizing a vision-language model. The authors report substantial enhanements over conventional text-based RAG for multimodal document retrieval and generation, identifying the significance of retaining document layout and visual information. VisRAG is specifically related to the proposed use of PDF and scanned-document datasets such as DocVQA. However, it does not concentrate on PII detection, privacy sanitization, or adaptive cloud-edge routing. The proposed research expands multimodal document RAG by adding a privacy-aware processing layer around retrieval and generation.

This paper [10] proposed SAGE, a two-stage synthetic-data approach for mitigating privacy risks in RAG. The method first extracts important attributes and generates synthetic retrieval data and subsequently applies iterative refinement to improve privacy protection. This work demonstrates that replacing private retrieval content with carefully generated synthetic information can reduce privacy risks while retaining useful contextual information. However, synthetic replacement can introduce information distortion and is different from the proposed approach, which seeks to preserve the original information locally and only sanitize sensitive components before selective cloud processing.

This research article [9] conducted one of the important empirical studies on privacy in RAG, The Good and The Bad: Exploring Privacy Issues in Retrieval-Augmented Generation. Their experiments establish that RAG systems can leak information from private retrieval databases throughout specifically designed attacks. At the same time, the research shows that RAG may minimize some forms of training-data leakage affiliated with LLMs. This work establishes an important theoretical basis for treating the retrieval database itself as a privacy-sensitive component rather than assuming that RAG automatically improves privacy. The proposed research extends this work by introducing a preventive architecture in which sensitive information is detected and sanitized before external cloud processing.

This article [11] proposed a vision-based multimodal RAG architecture for documents, showing that traditional text extraction can lose important visual and layout information. Their approach directly embeds document images using a vision-language model and uses them for retrieval and generation. The experiments show substantial end-to-end improvements over conventional text-based RAG pipelines. The research supports the proposed use of multimodal document processing and persuades the augmentation of scanned documents and images in the experimental design. However, privacy is outside the major scope of this work; COnsequently, the proposed research merges multimodal document retrieval with privacy detection, sanitization, and adaptive routing.

This paper [17] also explored the evaluation of RAG systems throughout a unified evaluation perspective. Their work determines that RAG cannot be properly assessed using only final-answer accuracy because retrieval quality and generation quality must be analyzed separately. Relevant dimensions take in retrieval relevance, answer correctness, faithfulness, and other generation-related procedures. This supports the proposed research methodology, which identifies not only answer quality but also Precision@K, Recall@K, MRR, NDCG@K,

faithfulness, response latency, communication cost, privacy leakage, and computational resource usage.

This study [13] represented a wide survey of multimodal retrieval for augmented generation and analyzed methods involving images, code, tables, graphs, and audio. The research emphasizes that multimodal retrieval can improve factuality, reasoning, interpretability, and robustness by feeding generative models with additional external information. It also recognizes the absence of a unified method for integrating different modalities into generation systems. This research provides the conceptual foundation for utilizing heterogeneous datasets in the proposed research. The proposed work expands the multimodal retrieval perspective by integrating privacy-aware preprocessing and adaptive local-cloud execution.

This research article [18] presented FLARE (Forward-Looking Active Retrieval Augmented Generation), which dynamically fetches information during the generation process rather than retrieving documents only once at the beginning. FLARE predicts oncoming content and prosecutes additional retrieval when confidence is low. This research establishes that retrieval can be dynamically controlled according to requires of generation. Although FLARE does not address privacy-aware local-cloud routing, its concept of adaptive retrieval decisions is conceptually relevant to the proposed adaptive routing mechanism. The proposed research extends adaptivity from retrieval decisions to privacy-aware execution-location decisions, determining whether processing should remain local or move to sanitized cloud processing.

This research paper [19] introduced MuRAG, one of the foundational multimodal RAG systems. MuRAG combines an external multimodal memory with a generative model and jointly processes image-text and text-only information. Experiments on WebQA and MultimodalQA demonstrated substantial improvements over competing approaches, establishing that retrieval-augmented generation can benefit from multimodal external knowledge rather than relying exclusively on textual retrieval. MuRAG accordingly feeds the foundational methodological basis for the multimodal RAG component of the proposed research. However, it does not address privacy-preserving local processing, PII detection, cloud sanitization, or adaptive cloud-edge routing. The proposed research extends this foundation toward a privacy-aware multimodal assistant architecture.

Regardless of these advances, the literature remains fragmented through several research directions. Multimodal RAG studies primarily focus on retrieval and generation quality; privacy studies primarily examine privacy attacks and information leakage; synthetic-data techniques focus on replacing private information; federated RAG focuses on distributed data and federated learning; and cloud-edge studies focus on adaptive inference and computational efficiency. These techniques provide valuable individual components but do not fully assimilate them into a single multimodal AI assistant architecture. In particular, there remains a require for a framework that combines local multimodal preprocessing, local RAG retrieval, privacy/PII detection, sensitive-information sanitization, adaptive local-versus-cloud routing, selective cloud generation, final privacy verification, and joint quantitative evaluation.

Material and Methods:

The proposed methodology is structured as a privacy-aware adaptive local–cloud multimodal Retrieval-Augmented Generation (PALC-MRAG) framework. It assimilate multimodal data processing, local retrieval, privacy detection, adaptive routing, selective cloud processing, privacy verification, and quantitative comparison into a single experimental pipeline. The architecture follows the standard that sensitive information should remain local whenever possible, while cloud processing should be used selectively when it provides adequate utility and only after sensitive information has been sanitized. This architecture constructs on the progression of multimodal RAG from image-text retrieval regarding broader

multimodal document understanding and adaptive retrieval systems. Current surveys examine that multimodal RAG must address heterogeneous representations, cross-modal retrieval, reasoning, efficiency, and robustness rather than performing all information as plain text [20].

Public Multimodal Datasets:

First, public multimodal datasets establish a controlled experimental knowledge base utilizing publicly available multimodal datasets. Wikipedia feeds textual knowledge, COCO assigns images and captions, arXiv feeds PDF/scientific documents, DocVQA provides scanned documents and question-answer pairs, Common Voice provides speech recordings and transcripts, and AudioCaps provides audio-caption pairs. The use of diverse datasets authorizes the proposed architecture to analyze whether the privacy-aware architecture behaves continually throughout various modalities rather than being optimized only for text. This multimodal framework is appropriate with current MRAG research, which examines text, images, audio, video, and multimodal documents as an important sources for retrieval-augmented generation [5]. For real time implementation used datasets are shown in table 2.

Table 2. Datasets used for real time implementation

Dataset	Source / version to pin	Modality	Planned sample	Selection criteria
Wikipedia	Wikimedia dumps; record exact dump date and language edition	Text	5,000–10,000 articles	English, non-empty, deduplicated; sample by fixed random seed
arXiv	arXiv public metadata and legally accessible PDFs; record download date	PDF/scientific documents	100–500 papers	Valid PDF, extractable text, broad subject sampling; remove duplicates
COCO	COCO 2017 images and captions/annotations	Images + text	500–1,000 images	Valid image and available annotation; stratify by category where possible
DocVQA	DocVQA dataset release; record release/configuration	Scanned/document images + QA	1,000–3,000 examples	Valid image/question/answer; avoid incomplete samples
Common Voice	Mozilla Common Voice; pin release number and language	Audio + transcript	500–2,000 clips	One language, valid audio, transcript present, duration/quality limits
AudioCaps	AudioCaps release; record split and audio availability	Audio + caption	500–1,000 clips	Available audio and caption; exclude broken links/files

Preprocessing:

Second, preprocessing prosecutes modality-specific preprocessing before information enters the local RAG system. Text is cleaned and normalized, PDF files are changed into structured textual or document representations, scanned documents endure OCR, images are resized and transformed into machine readable forms, and audio recordings are converted into text transcripts utilizing speech recognition such as Whisper. The process also entails cleaning duplicate or corrupted samples, normalization, feature extraction, encoding, chunking, and sampling. This procedure is significant because multimodal RAG cannot endorse that all input fact can be represented utilizing the same preprocessing procedure. Recent multimodal RAG

research particularly emphasizes the challenges of heterogeneous representations and the importance of preserving document structure, visual information, and cross-modal relationships [21].

Model and Retrieval Configuration:

For reproducibility, the implementation will pin exact model repository IDs, release tags or commit hashes, package versions, and configuration files. The following are proposed initial settings and must be recorded as actual configuration values before running the final experiments.

Table 3. Actual Configuration Value

Component	Candidate model/tool	Representation / dimensions	Initial setting
Text embeddings	BAAI/bge-base-en- v1.5 [22]	768-dimensional vectors	512-token chunks; 64-token overlap
Image embeddings	openai/clip-vit-base-patch32 [23]	512-dimensional image/text projected vectors	Model-native image preprocessing
Audio transcription	openai/whisper-small [24]	Transcript text; audio embeddings evaluated separately if selected	Pin model revision; record sample rate
Document OCR	PaddleOCR or Tesseract (pin version)	OCR text plus page/layout metadata	OCR locally; log confidence
Vector index	FAISS (pin package/version)	Cosine similarity after L2 normalization	Top-K=5; evaluate K in {3,5,10}
Text generation	Local model via Ollama/Hugging Face; cloud API named and version-pinned at run time	Generated answer plus retrieved evidence IDs	Fixed prompt and decoding settings
PII detection	Microsoft Presidio Analyzer/Anonymizer plus custom modality adapters	Entity labels, spans, confidence/risk features	Record recognizers, patterns, and thresholds

Embedding dimensions above are the expected dimensions for the named candidate checkpoints; they must be confirmed from the pinned model configuration at implementation time. Do not combine vectors from incompatible embedding spaces. Text and OCR text can share the text encoder; images use a vision-language encoder; audio is represented by local transcripts and, if a compatible audio encoder is implemented, audio embedding's. Cross-modal search will use a documented alignment strategy or separate modality-specific indexes followed by score normalization and rank fusion. Similarity will use cosine similarity on L2-normalized vectors. The primary experiment will use top-K=5, with K sensitivity analysis at 3 and 10. These are protocol choices, not experimentally optimized results.

Local Embedding's/RAG:

Local embedding's demonstrates the local retrieval layer of the proposed methodology. Instead of instantly transferring multimodal information to a cloud model, the preprocessed information is transformed into modality-appropriate embedding's. Text can be defined using models such as Nomic embedding's or BGE, images utilizing CLIP-style visual representations, audio using speech/audio representations, and documents using OCR

combined with visual or textual embedding's. The finding representations are saved in a local vector database such as ChromaDB or FAISS. When a user submits a question, the local retrieval component implements similarity search and selects the most related top-K contexts. This step is grounded in the fundamental RAG standard of retrieving external knowledge before generation. MuRAG established that expending the retrieval memory from text to images and text can substantially enhance multimodal question answering, while more developed multimodal RAG research has extended this concept to documents, audio, and other modalities [19].

Privacy Detection:

Privacy detection presents the main privacy-preservation mechanism**.** After retrieval, the system evaluates the query and relevant contextual information for sensitive information such as names, email addresses, telephone numbers, physical addresses, identification numbers, medical information, and other PII. A mechanism such as Microsoft Presidio or a customized privacy detector can be used to examine sensitive entities. The system assigns an adequate sensitivity level to the detected information and prosecute whether sanitization is required. This step is supported by previous research representing that RAG systems can leak information from retrieval databases and that multimodal RAG introduces additional privacy risks because sensitive information may be contained in images, documents, audio, or multimodal representations. Thus, privacy protection is cultivated as an explicit architectural component rather than an afterthought.

Adaptive Routing:

The adaptive routing component implements the adaptive routing mechanism**,** which is the central methodological contribution of the proposed architecture. Rather than using a fixed local-only or cloud-only strategy, a decision engine determines the sensitivity level and processing policy associated with the query and retrieved information. Low-risk information can remain in the local environment, because information requiring external model capabilities can be scrubbed and selectively transferred to the cloud. Consequently, the proposed architecture relaxes the conventional hypothesis that every request must utilize the same processing location: the execution environment becomes an adaptive decision based on privacy sensitivity and processing requirements**.** This thought is suitable with developed research on adaptive cloud-edge LLM inference, where execution can be vigorously selected according to privacy and utility considerations, while the proposed research extends the thought especially to multimodal RAG.

Local or Sanitized Cloud Processing:

The sanitized cloud processing contains two feasible execution paths. For low-risk requests**,** processing remains local and may utilize a local LLM, multimodal model, and retrieved RAG context. For requests that need additional cloud capabilities, the privacy module first masks, removes, or replaces sensitive information; only then is the sanitized information transmitted to a cloud LLM/API. This strategy therefore does not take over that cloud processing is inherently unacceptable. Instead, it follows a selective-cloud principle, in which cloud resources are utilized when they provide adequate utility while unnecessary exposure of sensitive information is reduced. This varies from conventional cloud RAG, where queries and retrieved contexts may be directly transmitted to remote services. Developed work on privacy-preserving cloud RAG and privacy-aware RAG establishes the significance of controlling what information is exposed to remote retrieval or generation services.

Response Generation:

This step creates the output response using either the local processing path or the sanitized cloud processing path. Depending on the input modality and application, the remark can include a document-related output/image, textual answer, audio response, or a multimodal

result. The retrieved proof is integrated into the generation process to enhance grounding and degrade dependence on the model's parametric knowledge. The use of multimodal retrieval is supported by the development of systems such as MuRAG and subsequent multimodal RAG frameworks, which demonstrate that external multimodal evidence can improve reasoning over heterogeneous information [19].

Privacy Verification:

The privacy verification fulfills post-processing privacy verification**.** The generated response and any information transmitted to the cloud are restrained for lasting sensitive entities. The system authenticates whether sanitization was successful and assesses privacy-related procedures such as privacy leakage rate, privacy protection rate, sensitive-data transmission, and PII detection performance. This measure is important because privacy cannot be certified primarily by involving a preprocessing filter; information may absorb in retrieved contexts, transformed representations, or generated outputs. Modern research on privacy attacks alongside multimodal RAG reinforces the need to evaluate leakage through the RAG pipeline rather than adopting that retrieval or multimodal processing is automatically private.

Quantitative Evaluation:

The quantitative evaluation assesses the proposed system quantitatively from several perspectives. Answer quality is studied utilizing Accuracy, F1, ROUGE/BLEU, faithfulness, and relevancy. Retrieval performance is examined utilizing Precision@K, Recall@K, MRR, and NDCG@K. Privacy performance is assessed using PII detection Precision/Recall/F1, sensitive-data transmission, privacy leakage rate, and privacy protection rate. Efficiency is studied through end-to-end latency, communication cost, and CPU/GPU/RAM usage. Finally, multimodal performance is analyzed using modality-specific accuracy/F1 and cross-modal retrieval steps. This multidimensional evaluation is consistent with developed multimodal RAG literature, which identifies that RAG systems should not be analyzed using final-answer accuracy alone; retrieval, generation, multimodal alignment, robustness, and efficiency must also be considered [5]. Metrics are measured on the held-out set and reported overall and by modality. The manuscript will not insert numerical resulted until the planned experiments have been executed.

Answer accuracy: exact match (EM) for tasks with reference answers:

$$EM = (1/N)\Sigma\, \mathrm{I}(\hat{y}_i = y_i)$$

For classification-style tasks, report precision, recall, and F1; for open-ended answers, use a documented human or model-assisted rubric with adjudication and report the rubric.

PII detection:

$$Precision = TP/(TP + FP)$$
$$Recall = TP/(TP + FN)$$
$$F1 = 2PR/(P + R)$$

Retrieval Precision@K = relevant retrieved items / K; Recall@K = relevant retrieved items / all relevant items;$\mathrm{MRR} = (1/N)\,\Sigma(1/\mathrm{rank}_i)$;nDCG@K = DCG@K/IDCG@K using graded relevance labels.

Privacy Leakage Rate (PLR): PLR = L/T, where L is the number of tested outputs and T is the number of evaluated test cases.

Sensitive-data Transmission Rate: SDR = cases with crude protected data sent to cloud.

Privacy Protection Rate: $PPR = 1 - PLR$ for the specified attack/test set.

Latency: end-to-end ceased time from query receipt to report median, final answer; standard deviation, mean, and 95th percentile. Decompose privacy scan, retrieval, preprocessing, cloud wait, and generation where instrumented.

Communication Cost: transmitted request + response bytes per query; total bytes and report average and cloud-call count. Do not compare token count with bytes unless the transformation method is specified.

Resource Use: peak and average CPU, GPU uses/ record sampling interval VRAM, and RAM and measurement tool.

Utility-Privacy Trade-off: draw retrieval metrics and answer-quality alongside PLR/SDR over a pre-declared threshold sweep; avoid failing them into a single score unless weights are verified in advance.

Statistical Analysis and Ablation:

For each report sample size, metric, median [IQR], mean ± standard deviation, and 95% confidence intervals where suitable. Use binary comparisons because entire systems answer the same test queries. For roughly normal paired differences, utilize a paired t-test; diversely use the Wilcoxon signed-rank test. For several baseline comparisons, employ Holm correction. Report effect sizes against p-values. Utilize bootstrap confidence intervals for PLR and retrieval measures where appropriate. Serve $p < .05$ as a conventional threshold while identifying uncertainty and effect size, not p-values alone.

Planned ablations: (A1) disable sanitization; (A2) relocate adaptive routing and utilize fixed hybrid routing; (A3) relocate indirect-disclosure checks; (A4) disable post-generation privacy verification; (A5) relate modality-specific retrieval alongside a single text-converted index; and (A6) vary risk thresholds and top-K. Disabling sanitization is limited to restrained synthetic test data and must never send real personal information to a cloud service.

Comparative Analysis:

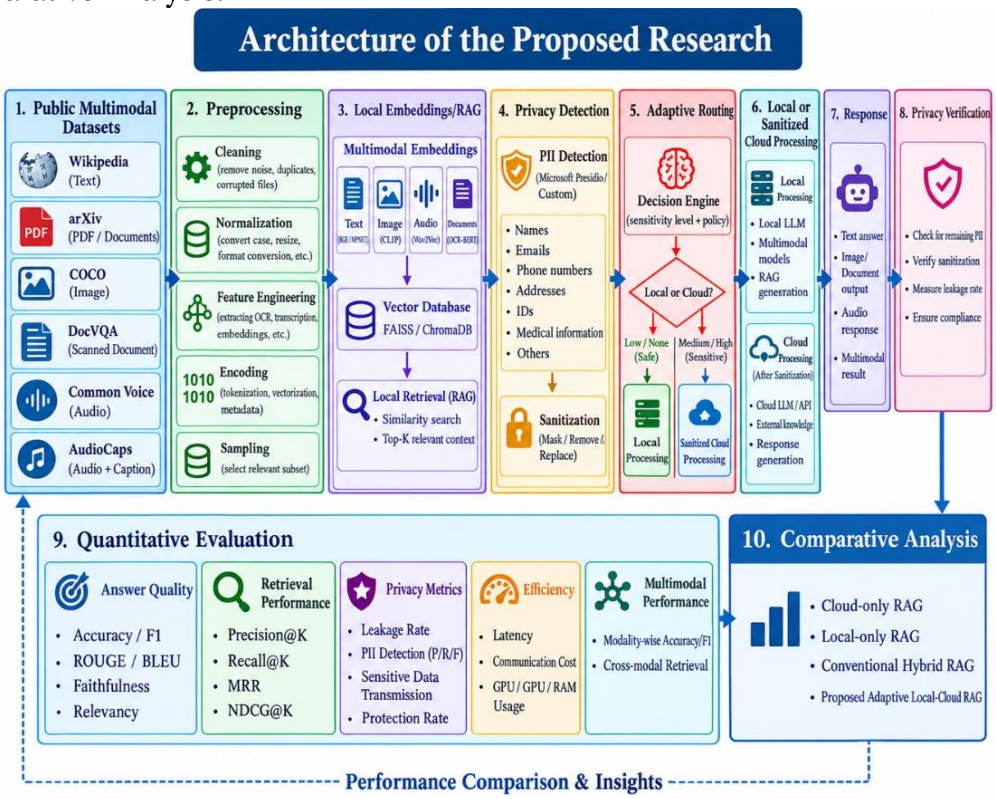

Figure 1. Architecture of Proposed Research

The comparative analysis pertains four experimental configurations**:** (1) cloud-only RAG, (2) local-only RAG, (3) and conventional hybrid RAG (4) the proposed adaptive local-cloud RAG**.** The aim is to determine whether adaptive routing can achieve a better balance between privacy and system utility than fixed processing techniques. The comparison considers accuracy, communication cost, retrieval quality, latency, privacy leakage, and

computational requisites. The investigation consequently tests the central assumption of the research: that selective cloud processing after privacy detection and sanitization can reduced unnecessary sensitive-data transmission while preventing acceptable retrieval performance and response quality.

The dashed suggestion path labelled "Performance Comparison & Insights" illustrates the final analytical loop. Findings from the quantitative evaluation and comparative experiments are described to analyze the privacy-performance trade-off and perform under which conditions local processing, cloud processing, or adaptive processing is most suitable. This makes the architecture an experimental method rather than just a software pipeline: each step generates limited outcomes that provide to the final comparison. Proposed research architecture is shown in figure 1.

Result and Discussion:

This research is expected to develop a functional privacy-preserving multimodal AI assistant combining local RAG with selective cloud processing. It is expected to certify that local processing can reduce unnecessary transmission of sensitive information across PDF, text, scanned-document, image, and audio data. The proposed privacy analyzer is predicted to examine and sanitize sensitive information before data is transferred to cloud services. The adaptive routing method is predicted to reduce privacy leakage while protecting acceptable response quality and retrieval accuracy. Using datasets such as DocVQA, COCO, Wikipedia, arXiv and AudioCaps, the research will assess performance throughout several modalities. Comparative experiments are predicted to expose differences among cloud-only, local-only, conventional hybrid, and proposed adaptive methodologies in accuracy, latency, and computational cost. The study is predicted to provide measurable identification of the privacy-performance trade-off in multimodal RAG systems. Ultimately, the results may provide a practical foundation for more privacy-aware, efficient, and trustworthy multimodal AI assistants. The research is expected to propose a Privacy-Aware Adaptive Local-Cloud Multimodal RAG (PALC-MRAG) framework for privacy-preserving AI assistance. It will propose an adaptive routing mechanism that prosecutes whether multimodal queries should be processed locally or selectively through sanitized cloud processing. The research will blend multimodal local retrieval, sensitive-information detection, data sanitization, and final privacy verification into a compacted pipeline. A proposed multimodal experimental dataset will be prepared from AudioCaps arXiv, COCO, Common Voice, DocVQA, and Wikipedia for predicting privacy and RAG performance. A comparative simulation will determine cloud-only, conventional hybrid, local-only, and adaptive local-cloud approaches utilizing retrieval, latency, privacy, accuracy, and computational metrics. The resulting architecture is predicted to allocate a practical research foundation for privacy-aware and efficient multimodal AI assistants in privacy-sensitive applications. The required result metrics of this research is shown in table 3.

Table 4. Required Result Metrics

Configuration	Answer quality	Retrieval	PLR (%)	Sensitive bytes/query	Latency (ms)	RAM/VRAM
B1 Cloud-only RAG	✓	✓	✓	✓	✓	✓
B2 Local-only RAG	✓	✓	✓	✓	✓	✓
B3 Conventional hybrid RAG	✓	✓	✓	✓	✓	✓
B4 PALC-MRAG	✓	✓	✓	✓	✓	✓

Practical Implications:

If verified, the architecture could help organizations settle which multimodal queries can persist on-device and which can benefit from a cloud model after sanitization. Potential use cases involve enterprise knowledge assistants, education documents, internal research repositories, and administrative document workflows. High-stakes healthcare or financial deployment would need additional domain verification, security assessment, legal review, and human regulation. The proposed experiments use public and synthetic data and cannot alone demonstrate compliance or safety for financial, real patient, or employment records.

Limitations:

Public datasets and synthetic PII scenarios may not describe the ambiguity, distributions, and contextual sensitivity of real private data.

PII detectors can produce false positives and false negatives; masking can remove context needed for accurate answers.

Embedding-based retrieval may expose semantic or membership information even when direct identifiers are removed.

Results will depend on pinned model versions, hardware, API behavior, network conditions, language, and modality.

Privacy leakage metrics cover defined attacks and seeded canaries; they cannot prove absence of all privacy leakage.

Cloud APIs may change, incur costs, impose rate limits, or provide limited reproducibility.

The planned study is a controlled prototype, not a production deployment or formal security proof.

Conclusion:

The research recommends a privacy-preserving multimodal AI assistant that blends RAG with selective cloud processing to regulate the performance, privacy, and computational objections associated with multimodal AI systems [7][25][26]. The proposed architecture is designed to process including text, PDF documents, images, heterogeneous data, audio, and scanned documents, while reducing unnecessary transmission of sensitive information to external cloud services [16][7]. By merging the local embedding's and retrieval, local multimodal preprocessing, data sanitization, privacy detection, adaptive routing, and privacy verification, the research aims to establish a practical approach for balancing privacy protection with AI performance [16]. The predicted findings are anticipated to identify that local processing can minimize the exposure and transmission of sensitive information, so long as selective cloud processing can feed access to additional computational and model capabilities when required [27]. The privacy analyzer and sanitization methodologies are expected to interrogate sensitive information before external processing, so long as the adaptive routing technique is predicted to select a suitable processing environment according to the sensitivity and processing requirements of each request [28]. Evaluation utilizing arXiv, COCO, Common Voice, DocVQA, and AudioCaps, Wikipedia will provide a multimodal experimental basis for imposing the proposed architecture [29][30][27]. Furthermore, comparative experimentations concerning conventional hybrid RAG, local-only RAG, cloud-only RAG, and the proposed adaptive local-cloud RAG are predicted to provide quantitative evidence regarding controversies in response quality, retrieval accuracy, privacy leakage, communication cost, latency, and computational resource utilization. The research is thus expected to contribute measurable evidence regarding the privacy-performance trade-off in multimodal RAG systems. Generally, the proposed research may provide a foundation for developing more privacy-aware, efficient, and trustworthy multimodal AI assistants, specifically for applications involving sensitive personal, organizational, or research information.

Future Work:

The proposed research will provide an experimental framework and simulation-based evaluation, real-world implementation and deployment will be necessary as a prominent direction for future work**.** A future implementation could recognize the proposed framework into an actual multimodal AI assistant executing on personal computers, mobile devices, edge servers, or organizational infrastructure. Such an implementation would allow the adaptive routing approach to be analyzed under realistic network conditions, workloads, hardware limitations, and user interactions. Future research could also explore larger and more disparate real-time multimodal queries, stronger local multimodal models, private datasets, and additional privacy-preserving techniques such as federated learning, secure computation, differential privacy, and encryption. Moreover, real-world deployment could analyze scalability, robustness against adversarial privacy attacks, cloud-service cost, energy consumption, and long-term privacy compliance. These evolutions would help move the proposed research from a restrained experimental environment to a fully deployable, implemented, and practically validated privacy-preserving multimodal AI assistant**.**

Recommendation:

The research paper "Privacy-Preserving Multimodal AI Assistant Using Local Retrieval-Augmented Generation and Selective Cloud Processing" is recommended for furthermore development because it addresses privacy risks in multimodal AI systems. The proposed PALC-MRAG architecture should be analyzed throughout experiments comparing cloud-based, hybrid processing, and local techniques. Particular attention should be given to latency, response accuracy, retrieval performance, privacy leakage, and data transmission. Future work should verify the architecture throughout real-world implementation and reproducible experiments to establish its effectiveness and practical applicability.

Acknowledgement: The author gratefully acknowledges the support and resources that promoted the completion of this study. The author also appreciates the helpful contributions of the researchers whose work informed this research.

Author's Contribution: The corresponding author was solely responsible for all aspects of this research, including data collection, conceptualization, methodology, investigation, analysis, and implementation, manuscript preparation, and validation. the author also conducted the literature review, interpreted the results, and finalized the complete manuscript.

Conflict of interest: There exists no conflict of interest for publishing this manuscript in IJIST.

References:

[1] B. Ni *et al.*, "Towards Trustworthy Retrieval Augmented Generation for Large Language Models: A Survey," *ACM Comput. Surv.*, vol. 58, no. 15, pp. 1–36, Feb. 2025, doi: 10.1145/3837074.

[2] B. Palanisamy, G. S. S. Chalapathi, V. Hassija, and R. Buyya, "Security and Privacy in Retrieval-Augmented Generation: Architectures, Threats, Defenses, and Future Directions for Building Trustworthy Systems," Jun. 2026, Accessed: Oct. 29, 2026. [Online]. Available: https://arxiv.org/pdf/2606.25533

[3] Y. Zhang *et al.*, "Privacy protection in RAG: A novel method and evaluation framework," *Inf. Process. Manag.*, vol. 63, no. 3, p. 104505, Apr. 2026, doi: 10.1016/J.IPM.2025.104505.

[4] G. Zhang, M. Tian, X. Luan, Y. Ma, and J. Liu, "Is External Database Protection Static in Retrieval-Augmented Generation? Rethinking Privacy Preservation under Dynamic Queries," vol. 14, no. 8, Jul. 2026, Accessed: Oct. 29, 2026. [Online]. Available: https://arxiv.org/pdf/2607.14811

[5] M. M. Abootorabi *et al.*, "Ask in Any Modality: A Comprehensive Survey on Multimodal Retrieval-Augmented Generation," *Proc. Annu. Meet. Assoc. Comput.*

Linguist., pp. 16776–16809, Feb. 2025, doi: 10.18653/v1/2025.findings-acl.861.

[6] J. Zhang *et al.*, "Beyond Text: Unveiling Privacy Vulnerabilities in Multi-modal Retrieval-Augmented Generation," May 2025, Accessed: Oct. 29, 2026. [Online]. Available: https://arxiv.org/pdf/2505.13957

[7] Y. Cheng, L. Zhang, J. Wang, M. Yuan, and Y. Yao, "RemoteRAG: A Privacy-Preserving LLM Cloud RAG Service," Dec. 2024, Accessed: Oct. 29, 2026. [Online]. Available: https://arxiv.org/pdf/2412.12775

[8] A. Chakraborty, C. Dahal, and V. Gupta, "Federated Retrieval-Augmented Generation: A Systematic Mapping Study," *EMNLP 2025 - 2025 Conf. Empir. Methods Nat. Lang. Process. Find. EMNLP 2025*, pp. 7362–7374, May 2025, doi: 10.18653/v1/2025.findings-emnlp.388.

[9] S. Zeng *et al.*, "The Good and The Bad: Exploring Privacy Issues in Retrieval-Augmented Generation (RAG)," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 4505–4524, Feb. 2024, doi: 10.18653/v1/2024.findings-acl.267.

[10] S. Zeng *et al.*, "Mitigating the Privacy Issues in Retrieval-Augmented Generation (RAG) via Pure Synthetic Data," *EMNLP 2025 - 2025 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, pp. 24527–24558, Jun. 2024, doi: 10.18653/v1/2025.emnlp-main.1247.

[11] S. Yu *et al.*, "VisRAG: Vision-based Retrieval-augmented Generation on Multi-modality Documents," *13th Int. Conf. Learn. Represent. ICLR 2025*, pp. 101219–101243, Oct. 2024, Accessed: Oct. 29, 2026. [Online]. Available: https://arxiv.org/pdf/2410.10594

[12] S. Zhao, Y. Yang, Z. Wang, Z. He, L. K. Qiu, and L. Qiu, "Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely," Sep. 2024, Accessed: Oct. 29, 2026. [Online]. Available: https://arxiv.org/pdf/2409.14924

[13] S. J. Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, "Retrieving Multimodal Information for Augmented Generation: A Survey," *Find. Assoc. Comput*, pp. 4736–4756, 2023, doi: 10.18653/v1/2023.findings-emnlp.314.

[14] J. Chen *et al.*, "Privacy-Preserving Reasoning with Knowledge-Distilled Parametric Retrieval Augmented Generation," Sep. 2025, Accessed: Oct. 29, 2026. [Online]. Available: https://arxiv.org/pdf/2509.01088

[15] A. Al-Lawati and S. Wang, "Do Multimodal RAG Systems Leak Data? A Comprehensive Evaluation of Membership Inference and Image Caption Retrieval Attacks," Jan. 2026, Accessed: Oct. 29, 2026. [Online]. Available: https://arxiv.org/pdf/2601.17644

[16] J. Zhan, H. Shen, Z. Lin, and T. He, "PRISM: Privacy-Aware Routing for Adaptive Cloud-Edge LLM Inference via Semantic Sketch Collaboration," Nov. 2025, Accessed: Oct. 29, 2026. [Online]. Available: https://arxiv.org/pdf/2511.22788

[17] H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, and Z. Liu, "Evaluation of Retrieval-Augmented Generation: A Survey," *Commun. Comput. Inf. Sci.*, vol. 2301, pp. 102–120, Jul. 2024, doi: 10.1007/978-981-96-1024-2_8.

[18] Z. Jiang *et al.*, "Active Retrieval Augmented Generation," *EMNLP 2023 - 2023 Conf. Empir. Methods Nat. Lang. Process. Proc.*, pp. 7969–7992, May 2023, doi: 10.18653/v1/2023.emnlp-main.495.

[19] W. Chen, H. Hu, X. Chen, P. Verga, and W. W. Cohen, "MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text," *Proc. 2022 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2022*, pp. 5558–5570, Oct. 2022, doi: 10.18653/v1/2022.emnlp-main.375.

[20] X. Zhang, X. Dong, Q. Wang, D. Zhang, and F. Cao, "A Survey of Multimodal Retrieval Augmented Generation: Architecture, Technologies, Applications, and Suggestions," pp. 93–98, Sep. 2026, doi: 10.23919/AIDMCPS00019.2026.00052.

[21] S. Gao *et al.*, "Scaling Beyond Context: A Survey of Multimodal Retrieval-Augmented Generation for Document Understanding," pp. 4458–4489, Oct. 2025, doi: 10.18653/v1/2026.acl-long.204.

[22] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, D. Lian, and J.-Y. Nie, "C-Pack: Packed Resources For General Chinese Embeddings," *arXiv*, vol. 1, Sep. 2023, doi: 10.1145/3626772.3657878.

[23] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," *Proc. Mach. Learn. Res.*, vol. 139, pp. 8748–8763, Feb. 2021, Accessed: Oct. 02, 2023. [Online]. Available: https://arxiv.org/abs/2103.00020v1

[24] I. S. Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, "Robust Speech Recognition via Large-Scale Weak Supervision," *Open AI*, 2024, [Online]. Available: https://cdn.openai.com/papers/whisper.pdf

[25] L. He, P. Tang, Y. Zhang, P. Zhou, and S. Su, "Mitigating privacy risks in Retrieval-Augmented Generation via locally private entity perturbation," *Inf. Process. Manag.*, vol. 62, no. 4, p. 104150, Jul. 2025, doi: 10.1016/J.IPM.2025.104150.

[26] A. Hemmat, M. Moqadas, A. Mamanpoosh, A. Rismanchian, and A. Fatemi, "VAGUE-Gate: Plug-and-Play Local-Privacy Shield for Retrieval-Augmented Generation," pp. 3715–3730, Mar. 2025, doi: 10.18653/V1/2025.IJCNLP-LONG.194.

[27] H. Jin and Y. Wu, "CE-CoLLM: Efficient and Adaptive Large Language Models Through Cloud-Edge Collaboration," *Proc. IEEE Int. Conf. Web Serv. ICWS*, no. 2025, pp. 316–323, Nov. 2024, doi: 10.1109/ICWS67624.2025.00046.

[28] Y. Gao, W. Luo, C. Wang, N. S. Ahmad, X. Wang, and P. Goh, "A privacy-preserving multi-user retrieval system for multimodal artificial intelligence," *Sci. Reports 2026 161*, vol. 16, no. 1, pp. 10348-, Feb. 2026, doi: 10.1038/s41598-026-40734-w.

[29] M. Mathew, D. Karatzas, and C. V. Jawahar, "DocVQA: A Dataset for VQA on Document Images," *Proc. - 2021 IEEE Winter Conf. Appl. Comput. Vision, WACV 2021*, pp. 2199–2208, Jul. 2020, doi: 10.1109/WACV48630.2021.00225.

[30] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, and M. Najork, "WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning," *SIGIR 2021 - Proc. 44th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 2443–2449, Mar. 2021, doi: 10.1145/3404835.3463257.

Copyright © by authors and 50Sea. This work is licensed under Creative Commons Attribution 4.0 International License.