Review

Not peer-reviewed version

Augmenting Large Language Models with External Data Sources: A Systematic Review of Methodologies, Performance Metrics, and Information Fidelity

Soham Mukherjee *, John Le *, Chau Nguyen, Thai Vu

Posted Date: 10 April 2026

doi: 10.20944/preprints202604.0717.v1

Keywords: large language models; Retrieval-Augmented Generation; fine-tuning; data fidelity; information retrieval; AI safety; factuality

 doi:10.20944/preprints202604.0717.v1

Article

Augmenting Large Language Models with External Data Sources: A Systematic Review of Methodologies, Performance Metrics, and Information Fidelity

Soham Mukherjee *, John Le *, Chau Nguyen and Thai Vu

Institute of Cybersecurity and Cryptology, School of Computing and Information Technology, University of Wollongong, Wollongong, NSW 2522, Australia

* Correspondence: sm765@uowmail.edu.au (S.M.); johnle@uow.edu.au (J.L.)

Abstract

Large Language Models (LLMs) have materialised as revolutionary tools across various domains, showcasing exceptional capabilities in natural language processing and generation. However, their reliance on static pre-training data limits their ability to access up-to-date and domain-specific information. The existing research often treats augmentation strategies in isolation, and limited efforts have been made to systematically compare them through the lens of information integrity. This review focuses specifically on Retrieval-Augmented Generation (RAG) and Fine-tuning, identifying them as the two dominant paradigms for integrating external knowledge: RAG for retrieval-based context injection and Fine-tuning for parametric knowledge adaptation. While existing surveys predominantly focus on performance metrics like accuracy or latency, this paper addresses the critical gap of *data fidelity*-the preservation of truthfulness, integrity, and fairness during augmentation. We systematically synthesise empirical findings from diverse methodologies to determine how each approach mitigates hallucinations and bias. By comparing the trade-offs between retrieval-based context injection and parametric knowledge adaptation, this survey brings unique value to readers by providing a structured taxonomy, a unified evaluation framework, and actionable insights to guide future research and practical deployment of robust, high-fidelity LLMs.

Keywords: large language models; Retrieval-Augmented Generation; fine-tuning; data fidelity; information retrieval; AI safety; factuality

1. Introduction

Large Language Models (LLMs), such as those detailed by Zhang et al. [1] and Sun et al. [2], signify a substantial leap in artificial intelligence. While these models demonstrate sophisticated capabilities in natural language processing (NLP), their reliance on static pre-training corpora introduces a significant "knowledge cut-off." This limitation renders LLMs incapable of processing real-time information, frequently leading to the generation of factually incorrect or outdated content-a phenomenon widely recognised as "hallucination" [3,4].

To address these limitations, two primary augmentation paradigms have emerged: Retrieval-Augmented Generation (RAG) and Fine-tuning. RAG systems, as described by Xue et al. [3] and Chauhan et al. [5], mitigate the knowledge cut-off by retrieving relevant external document chunks to condition the model's generation. Conversely, Fine-tuning involves continued training on domain-specific datasets to adjust internal parameters, deeply embedding specific knowledge or stylistic nuances [6]. While advanced variations exist-such as Dynamic RAG [4] and Knowledge-Graph enhanced pre-training [1]-current research largely evaluates these methods based on surface-level performance metrics rather than the integrity of information assimilation. Figure 1 provides a timeline illustrating the rapid evolution of these LLM augmentation methodologies.

 doi:10.20944/preprints202604.0717.v1

Figure 1. A timeline illustrating the rapid evolution of LLM augmentation methodologies, tracing the development from early parameter-heavy Fine-tuning to modern Agentic RAG and dynamic Hybrid systems.

A critical oversight in existing literature is the lack of comprehensive analysis through the lens of **Data Fidelity**. Current evaluations often reveal a misalignment between retrieval metrics and downstream performance [7], posing a risk of degrading public knowledge if AI outputs are uncritically accepted as a single source of truth [8].

 doi:10.20944/preprints202604.0717.v1

In this review, we explicitly define **Data Fidelity** as a multidimensional construct that encompasses three interactive layers: the properties of the external data sources (accuracy and curation), the interaction between the retrieval mechanism and the model (contextual relevance and grounding), and the properties of the final model outputs (truthfulness and structural coherence). We establish that Data Fidelity can be quantitatively and qualitatively measured through specific evaluation metrics, including retrieval accuracy, grounding correctness, hallucination rate, factual consistency, and bias/fairness indicators. This overarching concept rests on three core pillars:

1. **Truthfulness and Factuality:** The minimisation of stochastic hallucinations and adherence to verifiable facts.
2. **Integrity and Curation:** The resilience of the system against data poisoning [3] and context degradation.
3. **Representational Fairness:** The mitigation of algorithmic biases inherited from source data or amplified during augmentation [9].

Value to the Reader and Relation to Existing Surveys. Prior reviews typically focus on isolated aspects of the ecosystem, such as datasets [8], general data-augmentation methods [10], RAG-only architectures [11], or broad augmented LLM frameworks without a specific focus on integrity. In contrast, this review brings distinct value to the reader by offering:

- **A Novel Perspective:** Unifying disparate augmentation methods (RAG, Fine-Tuning, Hybrid) under the singular, critical umbrella of data fidelity and safety.
- **A Structured Taxonomy:** Clarifying the relationships, evolution, and architectural nuances of modern approaches through conceptual diagrams and structured categorisation.
- **Actionable Insights:** Providing a synthesised comparative framework to guide researchers and practitioners in selecting the optimal architecture based on their specific constraints (e.g., latency vs. provenance).

To this end, this survey addresses the following review questions, mapped to the corresponding sections of this paper:

- **RQ-A**: What are the main methodologies for integrating external data into LLMs? (Addressed in **Section 4**, providing a taxonomy of RAG, Fine-tuning, and Hybrid approaches).
- **RQ-B**: How do RAG, Fine-tuning, and Hybrid strategies compare in terms of fidelity and performance? (Addressed in **Section 6**, which synthesises comparative performance metrics).
- **RQ-C**: What open challenges and research gaps remain for developing trustworthy and high-fidelity augmented LLMs? (Addressed in **Section 7**, outlining future research categories).

The remainder of this paper is structured to follow a clear logical progression from foundational theory to practical application: **Section 2** establishes the *Conceptual Framework* motivating augmentation and defining fidelity threats. **Section 3** outlines the systematic review process. **Section 4** presents a *Method Taxonomy*, categorising the core architectures. **Section 5** outlines the *Evaluation Metrics* used to measure success. **Section 6** conducts a *Comparative Analysis*, synthesising the findings to evaluate strengths and weaknesses. Finally, **Section 7** outlines future directions and **Section 8** concludes our findings after synthesising our research insights.

2. Conceptual Framework and Background

The objective of this section is to establish the theoretical foundations of LLM augmentation and delineate the cross-cutting concerns that threaten data fidelity. Understanding these underlying vulnerabilities is crucial before evaluating the technical methods designed to address them.

2.1. The Need for Augmentation

Standard LLMs are limited by their training data cut-off and the stochastic nature of token generation. Hallucination which is the generation of plausible but incorrect information, remains a primary vulnerability [12]. Research indicates that an LLM's internal state may "know" when it is

 doi:10.20944/preprints202604.0717.v1

fabricating information, yet the token-by-token generation process can still result in a false output [12]. Augmentation strategies aim to ground these models in external reality. RAG achieves this via a non-parametric memory (external databases), while Fine-tuning relies on parametric updates. However, both methods introduce unique vectors for information loss and fidelity degradation. A high-level overview of these external data augmentation strategies is presented in Figure 2.

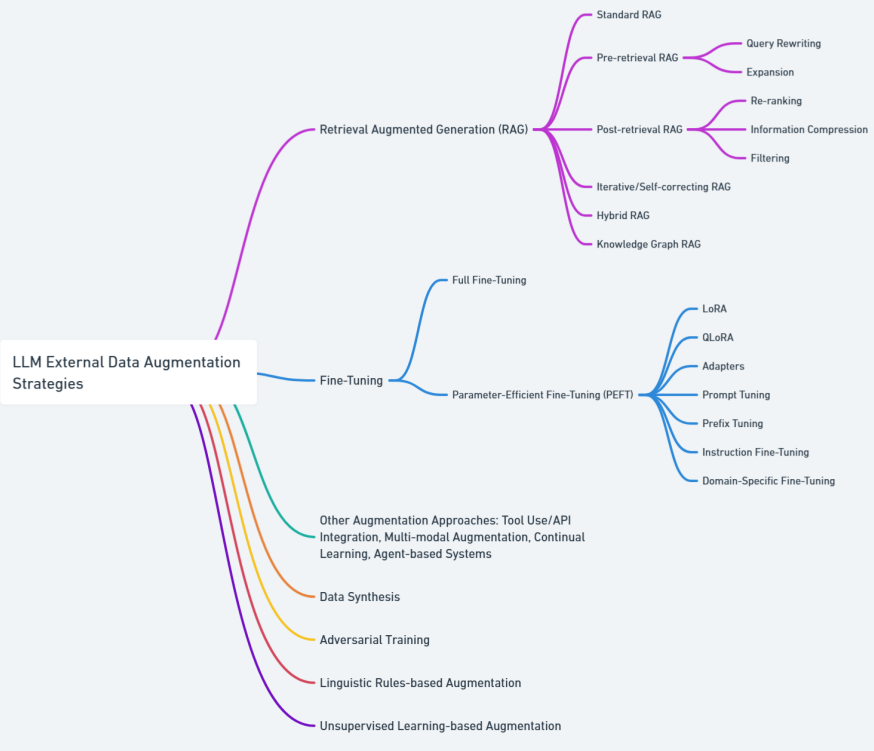

Figure 2. Overview of LLM External Data Augmentation Strategies.

2.2. Cross-Cutting Fidelity Concerns

Ensuring high fidelity requires addressing the intersection of security, privacy, and ethics.

2.2.1. Security: Vulnerabilities and Integrity

The operational security of LLMs directly impacts data integrity, with research identifying numerous vulnerabilities. Raj et al. [13] consider Prompt injection to be the primary threat, where malicious entities bypass safety filters by manipulating inputs. This is part of a broader set of risks catalogued in the OWASP "Top 10 for LLM Applications", which also includes threats such as insecure output handling and model denial of service [14].

In RAG systems, the retrieval pipeline represents a significant attack surface. The communication channel between the retriever and the generator is commonly vulnerable to data poisoning and "man in the middle" attacks, which will compromise the integrity and confidentiality of the data used by the LLM [15]. "Data poisoning" allows adversaries to insert malicious documents into the retrieval corpus, creating backdoors or skewing the generator's output. Furthermore, M'Lisha et al. [16] found that models frequently hallucinate details about zero-day vulnerabilities, highlighting the risk of relying on un-augmented models for security tasks. Security threats also extend to the physical level, LLMs are vulnerable to hardware-based attacks including side-channel attacks that leak information through power consumption and fault injection attacks that manipulate hardware to induce errors [17].

2.2.2. Privacy and Confidentiality

Protecting Personally Identifiable Information (PII) is a fundamental challenge in deploying LLMs due to the vastness of datasets which may contain such information inadvertently. Rathod et al. [18] highlight that LLMs can inadvertently leak training data verbatim. Beyond direct leakage, LLMs are

 doi:10.20944/preprints202604.0717.v1

vulnerable to inference attacks, which allow an attacker to infer sensitive attributes about individuals even if their data is not explicitly reproduced [18].

In RAG, privacy risks are exacerbated if the retriever accesses sensitive documents without proper access control, leading to "context leakage" where a user is presented with unauthorised data. Privacy-preserving techniques, such as Differential Privacy (DP) and Federated Learning (FL), are increasingly critical for high-fidelity Fine-tuning [19]. FL allows for model training on decentralised user data without centralising it, while DP adds statistical noise to limit the data shared during training.

2.2.3. Bias and Fairness

One of the most extensively studied issues in LLMs is their tendency to inherit and amplify societal biases from their training data [20]. This "representational fairness" is a core pillar of fidelity. Sharma et al. [21] found that conversational LLMs can create "generative echo chambers" by selectively reinforcing a user's existing beliefs and filtering out opposing views, which can increase opinion polarisation over time.

In RAG, bias can stem from the retrieval algorithm itself ("ranking bias"), prioritising popular but potentially biased sources over neutral ones ("popularity bias") or showing a preference for LLM-generated content ("source bias") [22]. Achieving true data fidelity requires mitigating these biases across both the model parameters and the retrieval corpus [23].

Summary of Conceptual Background: The fundamental limitation of static LLMs is their propensity for hallucination and knowledge decay. Augmenting these models is necessary, but introduces complex vectors for fidelity degradation. To maintain trust, augmentation architectures must be designed to actively resist data poisoning, preserve data privacy, and counteract inherent algorithmic bias.

3. Review Methodology

This study conducts a systematic literature review to map the landscape of LLM augmentation techniques. The review process follows the guidelines set out by PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) to ensure transparency and reproducibility.

3.1. Search Strategy and Selection

We searched major academic databases (ACM Digital Library, IEEE Xplore, Springer Link, Science Direct) and preprint archives (arXiv) using a combination of primary keywords ("Large Language Model", "Retrieval-Augmented Generation", "Fine-tuning") and secondary keywords ("Data Fidelity", "Hallucination", "Factuality").

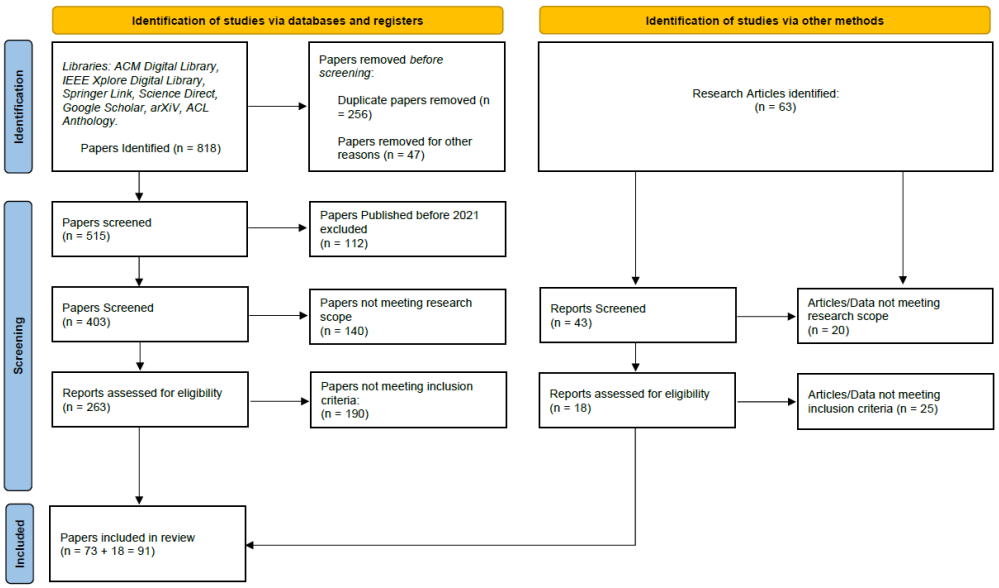

Figure 3. Paper Selection and Refinement Process represented in a PRISMA Flow

 doi:10.20944/preprints202604.0717.v1

3.2. Inclusion and Exclusion Criteria

Our initial search yielded 818 articles (Figure 3). To ensure the relevance and quality of the reviewed literature, we applied the following specific inclusion and exclusion criteria:

- **Inclusion Criteria:**
  - *Publication Date:* Articles published between January 1, 2021, and December 31, 2025, to capture the most recent advancements in LLM capabilities.
  - *Relevance:* Papers explicitly discussing "Retrieval-Augmented Generation," "Fine-Tuning," or "External Data Augmentation" in the context of Large Language Models.
  - *Focus on Fidelity:* Studies that include evaluations or discussions regarding data fidelity, factual consistency, hallucination mitigation, or information loss.
  - *Source Type:* Peer-reviewed journal articles and conference proceedings (e.g., ACL, NeurIPS, ICLR) to ensure academic rigor.
- **Exclusion Criteria:**
  - *Language:* Non-English publications.
  - *Scope:* Papers focusing solely on pre-training architectures without addressing external data integration.
  - *Format:* Opinion pieces, editorials, and non-peer-reviewed preprints (unless widely cited as seminal works in the field).
  - *Redundancy:* Duplicate studies or earlier versions of papers that have been subsequently published in journals.

After filtering for duplicates and relevance, 36 articles were selected for in-depth analysis. These were categorised into RAG, Fine-tuning, and Hybrid architectures to address the review questions (RQ-A, RQ-B, RQ-C).

4. Taxonomy of Augmentation Methodologies

The objective of this section is to categorise the primary methodologies for integrating external data into LLMs, providing a structured taxonomy (RQ-A). Rather than merely describing systems, we classify these methods by analysing them against clear comparison criteria: **accuracy** (precision of facts), **robustness** (resistance to noise), **latency** (computational efficiency), and **interpretability** (provenance tracking).

A structural comparison between Retrieval-Augmented Generation (RAG) and Fine-Tuning reveals divergent capabilities regarding knowledge management, hallucination mitigation, and data privacy. Fine-tuning deeply embeds domain-specific patterns into the parametric memory of the model, which is highly effective for adopting specific structural formats or stylistic nuances. However, it is inherently ill-suited for maintaining dynamic knowledge, as updating facts requires resource-intensive retraining. Conversely, RAG directly queries external databases at inference time, enabling real-time knowledge updates without altering the underlying model weights. Crucially, RAG frameworks are inherently less prone to hallucinations because generated responses are grounded in explicitly retrieved evidence. While fine-tuning can reduce hallucinations on familiar domain data, models remain highly susceptible to fabricating information when presented with unfamiliar inputs. Furthermore, RAG provides granular access control for privacy and ethical compliance through database-level security, whereas fine-tuning risks embedding sensitive or copyrighted material permanently into the model parameters. Table 1 highlights these fundamental conceptual differences.

 doi:10.20944/preprints202604.0717.v1

Table 1. Qualitative Comparison of RAG and Fine-Tuning Paradigms.

Feature	Retrieval-Augmented Generation (RAG)	Fine-Tuning
Knowledge Updates	Directly updates the retrieval knowledge base, ensuring information remains current without retraining. Highly suitable for dynamic data.	Requires continuous retraining to update knowledge. Not suitable for adding new knowledge or scenarios requiring rapid iteration.
Reducing Hallucinations	Inherently less prone to hallucinations since each answer is grounded in verifiable retrieved evidence.	Reduces hallucinations on specific domain data but may still exhibit severe hallucinations when faced with unfamiliar input.
Ethical and Privacy	Privacy concerns arise primarily from storing and retrieving text from external databases, which can be secured via access controls.	Ethical and privacy concerns arise from permanently embedding sensitive content directly into the model's training data parameters.

We categorise methodologies based on their architectural approach to fidelity: Retrieval-Based (RAG), Parameter-Based (Fine-tuning), and Hybrid/Agentic approaches. Figure 4 illustrates the conceptual relationships and overlaps between these dominant paradigms.

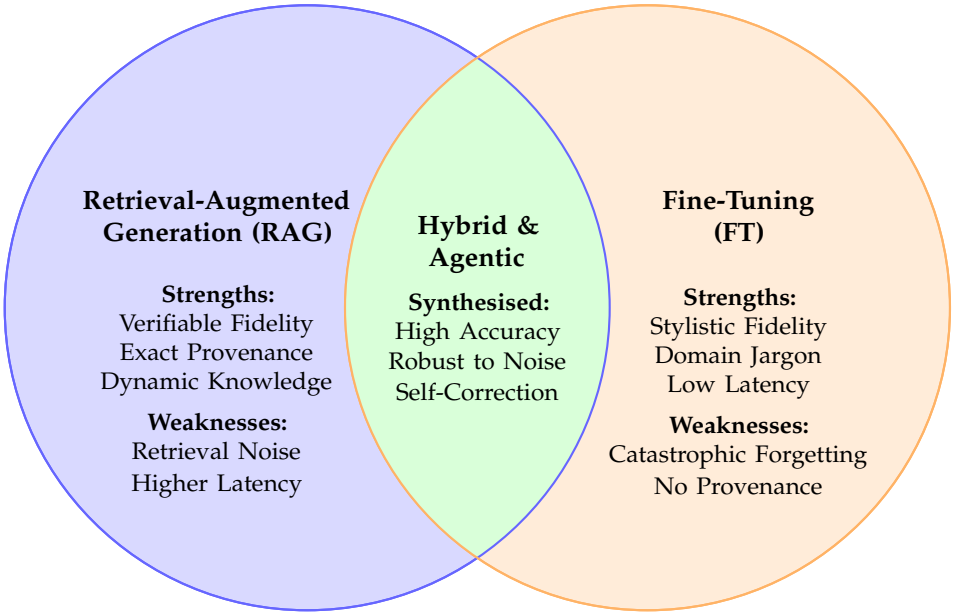

Figure 4. A conceptual Venn diagram demonstrating the overlapping relationships, shared benefits, and distinct capabilities of RAG, Fine-tuning, and Hybrid augmentation approaches.

4.1. Retrieval-Augmented Generation (RAG)

RAG integrates external data by retrieving relevant document chunks to condition the LLM's generation, akin to an "open-book" exam [11]. This process allows the model to retrieve current factual data to substantiate its responses, thereby enhancing accuracy and interpretability.

4.1.1. Naive RAG

The "Naive" approach is the most basic implementation. The system performs a simple semantic search (typically cosine similarity) to retrieve top-k documents from a vector database and concatenates them with the user prompt [24]. While easy to implement and highly interpretable, it suffers from low **Contextual Robustness**. It can struggle with getting the retrieval step right, often pulling in text that seems related on the surface but isnt actually helpful. Another issue is the "lost in the middle" phenomenon, where the LLM tends to ignore information buried in long contexts, leading to fidelity

 doi:10.20944/preprints202604.0717.v1

degradation [25]. However, practical implementations in enterprise environments demonstrate that combining RAG with automated extract-transform-load (ETL) processes is essential to maintain data currency [26].

4.1.2. Advanced RAG

To overcome the limitations of naive RAG, advanced techniques aim to improve the robustness and fidelity of the retrieved context:

- **Query Transformation and Decomposition:** Techniques like *Query Decomposition* break complex questions into sub-queries. The system finds documents for each sub-question and pieces the information together. *HyDE* (Hypothetical Document Embeddings) takes a different approach by asking the LLM to generate a theoretical answer to guide retrieval, ensuring the semantic search aligns with the answer space rather than just the question space [27].
- **Re-Ranking:** A second-stage process using cross-encoders to score documents based on true relevance. *RankRAG* [28] demonstrates that re-ranking significantly reduces noise. Unlike bi-encoders used in the initial search, cross-encoders process the query and document together, allowing for a deeper analysis of relevance. This directly improves the factual accuracy of the generation by filtering out irrelevant "distractors."
- **Knowledge Fusion:** Methods like *RAG-Fusion* [29] generate multiple query perspectives and fuse the results using Reciprocal Rank Fusion (RRF). This provides a richer context by combining diverse sets of retrieved documents. However, this negatively impacts the system's **latency** due to the multiple retrieval operations required.

4.2. Fine-Tuning Strategies

Fine-tuning modifies the model's internal weights to internalise knowledge or adapt behaviour. This involves continued training on domain-specific datasets, fundamentally altering the parametric memory of the model.

4.2.1. Knowledge Injection and Domain Adaptation

Approaches like *Synthetic Knowledge Ingestion* (SKI) [27] generate high-quality synthetic datasets to inject facts into the model. Strategies include fine-grained synthesis and interleaved generation to prepare data optimised for supervised fine-tuning. While highly effective for improving latency (as it removes the retrieval bottleneck), Fine-tuning suffers from "catastrophic forgetting"-the degradation of previously learned knowledge-posing a severe risk to **Integrity Fidelity** and offering near-zero **interpretability**.

4.2.2. Parameter-Efficient Fine-Tuning (PEFT)

To address efficiency and forgetting issues, PEFT methods like **LoRA** (Low-Rank Adaptation) freeze pre-trained weights and inject trainable rank decomposition matrices [30]. This reduces the number of trainable parameters by up to 10,000x while matching full fine-tuning performance. *RAFT* (Reward Ranked Fine-tuning) [31] aligns the model to prefer factual responses by training on "winning" samples ranked by a reward model, effectively steering the model towards **Truthfulness Fidelity** without the complexity of full reinforcement learning pipelines. Table 2 outlines key literature, applications, and limitations associated with these fine-tuning and PEFT methodologies.

 doi:10.20944/preprints202604.0717.v1

Table 2. Literature on Fine-tuning and PEFT: Methodology and Applications.

Paper	Methodology	Application / Benchmark	Key Results	Limitations & Future Scope
LoRA [30]	Freezes pre-trained weights, injects trainable low-rank matrices.	GPT-3 175B on RoBERTa/DeBERTa tasks.	Reduces trainable parameters by 10,000x, matches FFT performance.	Rank r is a sensitive hyperparameter.
RAFT [31]	Rejection sampling FT via reward model ranking.	Generative alignment on LLaMA-7B.	More stable than PPO-based RLHF for alignment tasks.	Overhead of generating k samples during training.
LoFiT [32]	Fine-tunes sparse subset (3-10%) of attention heads.	TruthfulQA on Llama2-7B.	Improved TruthfulQA accuracy from 62.2% to 74.4%.	Locating optimal "locus" of heads is non-trivial.
Multi-Fidelity FT [33]	Sequential training: low-fidelity then high-fidelity data.	Noise robustness benchmarks.	Outperforms naive mixing, leverages noisy data effectively.	Risk of catastrophic forgetting of initial stage.

4.2.3. Continual Learning with New Data

Continual Learning (CL) addresses the need to update models with new external information without complete retraining. [34] demonstrates that "Replay to Remember" mechanisms, coupled with LoRA, can stabilise performance and facilitate partial recovery of domain-specific knowledge when exposed to streaming data. However, despite advancements, current CL techniques do not completely eradicate catastrophic forgetting, especially when new data streams are highly dissimilar to previously learned domains.

4.3. Hybrid, Agentic, and Structured Approaches

The field is moving towards systems that combine the strengths of both paradigms to optimise across all comparison criteria.

4.3.1. Knowledge Graph Integration

Graph-enhanced LLMs, such as *GRAG* [35], retrieve structured subgraphs rather than text chunks. By linearising these graphs into text, they preserve relational information that flat text retrieval often loses, improving **Structural Fidelity** and robustness. *ERNIE* [1] and *KnowBERT* [36] integrate knowledge graph embeddings directly into the LLM architecture using techniques like TransE to fuse entity info.

4.3.2. Agentic Frameworks and Plugins

Agentic systems, such as *AgentFusion* [37], employ autonomous agents to handle retrieval, validation, and generation separately. *Tool-using agents* can query APIs or SQL databases dynamically. Frameworks like *MOYA* [38] employ multi-agent systems for CloudOps, where distinct agents handle security scanning and reporting. These systems offer high adaptability but introduce new security risks regarding autonomous execution and data access controls. The idea of an LLM acting as an intelligent agent is often achieved through "Reasoning and Acting" (ReAct) loops, where the model decides on its own when to use a function or call an API. Key contributions and limitations of recent literature concerning knowledge graphs, structured data, and agentic RAG frameworks are summarised in Table 3.

 doi:10.20944/preprints202604.0717.v1

Table 3. Literature on Knowledge Graph, Structured Data, and Agentic RAG.

Paper	Methodology	Application / Benchmark	Key Results	Limitations & Future Scope
ERNIE [1]	Pre-training with KGs, fuses entity info via TransE.	Entity Typing, Relation Classification.	Significant improvements on knowledge-driven tasks.	Static knowledge injection requires costly retraining.
GRAG [35]	Retrieves *k*-hop ego-graphs, "soft pruning".	Graph QA (WebQSP).	Outperforms standard RAG by preserving topological structure.	Subgraph retrieval is computationally intensive.
BIORAG [6]	Domain-Specific RAG, MeSH hierarchy refinement.	Biomedical QA.	73-90% accuracy on medical datasets.	Highly domain-coupled maintenance.
AgentFusion [37]	Multi-agent collaboration for retrieval/validation.	Technical documentation.	Significantly reduces hallucinations in non-English contexts.	High complexity in agent orchestration.

Summary of Method Taxonomy: Retrieval-based methods (RAG) excel in grounding outputs and interpretability but are susceptible to retrieval noise and high latency. Parameter-based methods (Fine-tuning) deeply embed domain logic and offer excellent latency, but struggle with rapid knowledge decay, lack provenance, and suffer from catastrophic forgetting. Hybrid architectures aim to balance these trade-offs by using retrieval for factual grounding and fine-tuning for domain-specific reasoning and stylistic alignment.

5. Evaluation Metrics for Fidelity

The objective of this section is to outline how the theoretical concept of data fidelity-as defined in the introduction-is practically and quantitatively measured across the distinct stages of the augmentation pipeline. Evaluating augmented LLMs requires assessing performance at three distinct levels: the retriever, the generator, and the end-to-end system.

5.1. Retrieval-Level Metrics

Fidelity begins with retrieval. Key metrics include **Hit Rate** (presence of ground truth in top-k) and **Mean Reciprocal Rank (MRR)**. High retrieval accuracy is essential to prevent "hallucination due to omission" [39]. If the retriever fails to fetch the relevant documents, even the most advanced LLM will struggle to produce a correct response, thereby severing the foundation of data fidelity.

5.2. Generation-Level Metrics

Metrics here directly quantify the **Truthfulness** pillar of data fidelity by evaluating the output against the retrieved context. **Factual Consistency** measures whether the generated claim is supported by the source, often using automated natural language inference (NLI) models [40], serving as a direct proxy for the *hallucination rate*. **Grounding Correctness** evaluates how well the generated text cites the provided sources. Traditional NLP metrics like BLEU and ROUGE are often insufficient as they measure n-gram overlap but fail to capture semantic correctness or factual hallucinations.

5.3. End-to-End Evaluation

Downstream performance is measured via **Exact Match (EM)** and **F1-Score**. However, Salemi et al. [7] note a misalignment between retrieval scores and generation quality, suggesting that simply fetching the "right" documents is not enough. The information must also be presented in a way that is usable by the LLM, necessitating holistic human evaluation and **bias/fairness indicators** to capture nuances of tone, safety, and representational fairness that automated metrics miss.

5.4. Computational Efficiency of Evaluation

A critical challenge in evaluating the data fidelity of RAG frameworks is the immense computational overhead required for End-to-End (E2E) evaluation. Traditional E2E evaluation processes feed the entirety of the retrieved document list into the LLM simultaneously, resulting in quadratic scaling

 doi:10.20944/preprints202604.0717.v1

of computational costs relative to input length. As detailed in Table 4, benchmarking standard RAG architectures using E2E methods on datasets like Natural Questions (NQ) and FEVER can consume between 46 GB and 75 GB of GPU memory and require over 3,000 seconds of runtime. Advanced evaluation methodologies, such as eRAG [7], mitigate this bottleneck by assessing documents individually. This document-level approach not only achieves higher correlation with true downstream performance but also reduces GPU memory consumption by up to 50 times (requiring only 1.5 GB uniformly) and accelerates runtime by a factor of 1.2 to 3.2. Efficient evaluation metrics are therefore paramount for the rapid iteration and deployment of secure augmented language models.

Table 4. Computational Efficiency: End-to-End (E2E) vs. Document-Level (eRAG) Evaluation.

Dataset	Runtime E2E (s)	Runtime eRAG (s)	Memory E2E (GB)	Memory eRAG-Doc (GB)
Natural Qs (NQ)	918	351	75.0	1.5
TriviaQA	1819	686	46.2	1.5
HotpotQA	1844	712	52.4	1.5
FEVER	3395	1044	66.5	1.5
Wizard of Wiki	912	740	47.9	1.5

Summary of Evaluation Metrics: Evaluating an augmented LLM requires a multi-tiered approach. Excellent retrieval metrics (MRR) do not guarantee high fidelity if generation metrics (Factual Consistency) reveal that the model ignores the retrieved context to hallucinate answers. Furthermore, shifting to document-level evaluation provides significant computational efficiency improvements over standard end-to-end approaches.

6. Comparative Analysis and Discussion

The objective of this section is to systematically synthesise the literature, moving beyond individual study summaries to derive actionable insights regarding the strengths, limitations, and optimal use cases for each augmentation paradigm. This directly addresses **RQ-B**.

Based on the aggregated performance metrics across various RAG and Fine-tuning models, we analyse the trade-offs between these paradigms across our core comparison criteria: accuracy, robustness, latency, and interpretability.

Table 5. Comparative Synthesis of LLM Augmentation Paradigms based on Data Fidelity Criteria.

Paradigm	Accuracy & Factual Fidelity	Robustness (to noise/decay)	Latency & Cost	Interpretability & Provenance
Naive RAG	High for retrieved facts, prone to hallucination if retrieval fails.	Low, highly susceptible to irrelevant "distractor" documents.	Medium, depends on vector search speed.	**High**, directly traces output to source chunks.
Advanced RAG (e.g., RankRAG)	**Very High**, re-ranking filters noise before generation.	High, robust against poorly phrased queries via query expansion.	High, re-ranking and multi-queries add significant delay.	**High**, maintains clear provenance to filtered sources.
Fine-Tuning (PEFT)	Medium, good for stylistic alignment but struggles with novel facts.	Low to Medium, prone to catastrophic forgetting over time.	**Low**, no retrieval step required at inference.	Low, knowledge is parametrically embedded (black-box).
Hybrid / Agentic	**Very High**, combines parametric reasoning with external facts.	**Very High**, agents can verify and self-correct retrieved data.	Very High, multiple LLM calls and tool uses required.	Medium to High, execution traces provide some transparency.

6.1. Accuracy vs. Factual Fidelity

Accuracy stands as a fundamental metric for evaluating the overall correctness of the generated responses. As synthesised in Table 5, RAG systems achieve **Verifiable Fidelity** by anchoring generation

 doi:10.20944/preprints202604.0717.v1

to retrieved evidence, yielding high interpretability. For instance, TC-RAG achieves high accuracy (83.15% on MMCU-Medical) by leveraging specialised retrieval (Table 6). This high accuracy demonstrates that domain-specific indexing is a prerequisite for high information fidelity in specialised fields, general-purpose retrieval often fails to capture the necessary semantic nuance.

In contrast, Fine-tuning achieves **Stylistic Fidelity**-internalising domain jargon-but risks lower **Factual Fidelity** on novel facts due to hallucination on unseen data. While fine-tuned models like T5-Large can achieve high accuracy on static benchmarks (0.91 on BoolQ), fine-tuning on new facts can often increase the tendency to hallucinate on adjacent topics. This suggests a core trade-off: achieving high accuracy on a specific set of newly learned facts via fine-tuning may come at the cost of degrading the model's broader information robustness.

While fine-tuning is traditionally viewed as a mechanism for knowledge injection, empirical evidence from Gekhman et al. [41] utilising the PaLM 2-S architecture suggests that acquiring novel factual knowledge via supervised fine-tuning actively degrades a model's pre-existing factual fidelity. When the PaLM 2-S model is fine-tuned on facts categorised as "Unknown" (information entirely absent from its pre-training), it exhibits severe overfitting. As demonstrated in Table 7, extending fine-tuning to convergence on unknown datasets ($D_{Unknown}$) results in catastrophic performance drops on the model's previously "HighlyKnown" facts, falling from 95.6% accuracy at early stopping to merely 55.8% at convergence. This indicates that attempting to force an LLM to internalise new facts linearly increases its tendency to hallucinate regarding its established parametric knowledge. Therefore, fine-tuning should primarily be utilised to expose and structure pre-existing knowledge rather than to inject entirely new factual databases.

Table 6. Methodological Evaluation of Specialized RAG Architectures

Framework	Target Domain	Evaluation Dataset	Evaluation Metric	Reported Performance
TC-RAG [42]	Medical QA	MMCU-Medical	Exact Match (EM)	83.15%
R2AG [43]	Open-domain QA	NQ-10	Exact Match (EM)	69.30%
VUL-RAG [44]	Software Security	PairVul	Precision	61.00%

Table 7. Impact of Fine-Tuning on Pre-existing Knowledge Fidelity (PaLM 2-S Architecture) [41]

Training Dataset Variant	HighlyKnown (Test)	MaybeKnown (Test)	WeaklyKnown (Test)	Unknown (Test)
Early Stopping (Optimal Generalisation)				
Trained on $D_{HighlyKnown}$	98.7%	60.1%	9.0%	0.6%
Trained on $D_{Unknown}$	95.6%	52.9%	6.5%	0.6%
Convergence (Overfitting on New Knowledge)				
Trained on $D_{HighlyKnown}$	98.4%	58.8%	8.5%	0.7%
Trained on $D_{Unknown}$	55.8%	36.6%	12.2%	3.2%

Table 8. Trade-offs Between RAG, Fine-Tuning, and Hybrid Paradigms (POPQA Dataset)

Model Architecture	RAG Only (Ideal)	Fine-Tuning (PEFT) Only	Hybrid Approach
StableLM2 (1.6B)	0.761	0.217	0.821
Llama3 (8B)	0.813	0.569	0.833

As shown in Table 8, Hybrid approaches often outperform individual strategies. On the Llama3 8B model, RAG alone achieves 0.813 accuracy, while Fine-tuning alone achieves only 0.569. The Hybrid approach improves this to 0.833, highlighting that RAG is the primary driver of factual accuracy for novel data, whereas fine-tuning alone often fails to effectively internalise new facts.

 doi:10.20944/preprints202604.0717.v1

6.2. Contextual Robustness and Recall

RAG offers superior recall for open-domain queries but faces challenges with fragmentation (context loss). Fine-tuning generally has lower recall for new facts but higher coherence. Table 9 illustrates that methods like *RankRAG* significantly boost recall by refining the retrieval stage, increasing robustness against noisy raw data.

Traditional RAG systems often require retrieving hundreds of short passages to achieve high recall (e.g., over 90%), which exacerbates context fragmentation. By contrast, LONGRAG demonstrates remarkable efficiency by processing Wikipedia into much larger 4K-token units, achieving a strong Answer Recall of 71.7% on NQ by retrieving *just a single document* (Recall@1). By retrieving fewer but longer, contextually complete documents, it effectively mitigates the "lost in the middle" problem while reducing the burden on the retriever. In contrast, Fine-tuning typically struggles to inject new knowledge and consistently underperforms RAG in recall-heavy tasks, making it a risky strategy for applications in rapidly evolving domains.

Table 9. Retrieval Completeness (Recall@k)

Model / Method	Dataset	Metric	Score (%)
LONGRAG [25]	Natural Questions (NQ)	Answer Recall@1	71.7
LONGRAG [25]	HotpotQA	Answer Recall@2	72.5
RankRAG (8B) [28]	TriviaQA	Recall@5	93.2
RankRAG (8B) [28]	TriviaQA	Recall@10	95.4

6.3. Interpretability and Safety (Exact Match)

To systematically assess information fidelity, it is necessary to benchmark the baseline hallucination propensity of foundation models across diverse generative tasks. Table 10 presents hallucination evaluation results utilising the TrustLLM framework [45] across multiple-choice (MC), open-ended question-answering (QA), knowledge-grounded dialogue (KGD), and text summarisation (SUM) tasks. The data reveals that state-of-the-art models exhibit highly task-dependent fidelity. For instance, GPT-4 achieves superior performance in objective tasks, scoring 0.835 in MC and 0.760 in summarisation, yet it struggles significantly with knowledge-grounded dialogue, dropping to an accuracy of 0.150. Conversely, models optimised via advanced alignment training, such as ChatGLM2, demonstrate robust performance in dialogue (0.500) and QA (0.600). The widespread failure of baseline LLMs to consistently surpass 0.600 accuracy across these generation tasks underscores the absolute necessity of external augmentation techniques like RAG to ensure high-fidelity outputs.

For precision-critical tasks, RAG frameworks demonstrate superior **Interpretability**, allowing for atomic verification of claims against source documents. Table 11 shows that advanced RAG methods like RankRAG achieve high Exact Match scores (up to 0.829 on TriviaQA), significantly outperforming standard baselines in precision-critical scenarios.

While RAG systems allow for verification, Fine-tuning lacks this traceability, once knowledge is encoded, the link to the source is severed. However, RAG introduces **Input Fidelity Risks**, where the system is susceptible to retrieving 'poisoned' or irrelevant contexts that immediately degrade the output, a vulnerability less prevalent in rigorously curated fine-tuning datasets.

 doi:10.20944/preprints202604.0717.v1

Table 10. Comprehensive Hallucination Evaluation Across Generation Tasks (Accuracy).

Model Architecture	Multi-Choice (MC)	Open QA	Dialogue (KGD)	Summarisation
GPT-4	0.835	0.320	0.150	0.760
ChatGPT (GPT-3.5)	0.557	0.500	0.430	0.630
ChatGLM2	0.557	0.600	0.500	0.510
Llama2-70B	0.256	0.370	0.440	0.540
Mistral-7B	0.412	0.480	0.450	0.490
Vicuna-33B	0.412	0.410	0.420	0.450

Table 11. Performance on Precision-Critical QA

Model / Method	Base LLM	Dataset	Metric	Score
R2AG [43]	LLaMA-2-7B	HotpotQA	Accuracy	0.667
R2AG (w/ RAFT) [43]	LLaMA-2-7B	HotpotQA	Accuracy	0.735
RankRAG [28]	LLaMA3-8B	TriviaQA	Exact Match (EM)	0.829

6.4. When to Use What?

Synthesising the comparative performance across these criteria reveals distinct operational domains for each paradigm:

- **Use RAG when:** Data changes frequently (News, Stock prices), provenance, interpretability, and citation are required, the knowledge base is vast (Web-scale). RAG is essential for dynamic, "knowledge-intensive" tasks where output must be strictly verified.
- **Use Fine-Tuning when:** The domain is static (Medical terminology, Legal syntax), low latency is critical (eliminating retrieval steps), the goal is stylistic adaptation or instruction following. Fine-tuning offers superior stylistic robustness and latency, but poor interpretability.
- **Use Hybrid when:** High precision is required on complex, domain-specific tasks that demand both deep structural understanding (FT) and up-to-date facts (RAG).

Summary of Comparative Analysis: The literature demonstrates that RAG is the primary driver of factual accuracy and interpretability for novel data, while fine-tuning alone fails to effectively internalise new facts without suffering from hallucination. Hybrid approaches reliably yield the highest performance by offsetting the limitations of each individual paradigm.

7. Future Research Directions

This section addresses RQ-C: *What open challenges remain?*

We identify critical gaps that serve as a roadmap for future research into high-fidelity augmentation.

7.1. Quantifying Data Fidelity

There is a lack of standardised metrics for "fidelity." Future research should develop unified frameworks that quantify the three pillars (truthfulness, integrity, fairness) simultaneously, rather than treating them as separate optimisation targets. Metrics like FactScore and TruLens are a step in the right direction, but broader adoption is needed.

7.2. Continual Learning Without Forgetting

While PEFT mitigates catastrophic forgetting, it does not eliminate it. Developing robust Continual Learning (CL) algorithms that can integrate streaming data updates without degrading existing knowledge remains a "holy grail" for parametric augmentation. Future work should explore adaptive "switch" mechanisms that dynamically determine whether to retrieve external context or rely on parametric knowledge.

 doi:10.20944/preprints202604.0717.v1

7.3. Privacy-Preserving Augmentation

As RAG systems are deployed in enterprise environments, "Private RAG" becomes critical. Future work must explore homomorphic encryption and secure enclaves to allow retrieval over sensitive data without exposing the raw content to the LLM provider or the retrieval index. This is essential to ensure that augmenting models with sensitive external data does not lead to inadvertent data leakage.

7.4. Agentic Reliability and Safety

As systems move towards autonomous agents (Agentic RAG), ensuring the reliability of tool use is paramount. Research is needed into "guardrails" that prevent agents from executing harmful API calls or accessing restricted data during the augmentation loop.

8. Conclusion

This systematic review mapped the landscape of methodologies for augmenting large language models with external data, specifically contrasting Retrieval-Augmented Generation (RAG) and Fine-tuning through the lens of data fidelity. Our analysis identified a distinct dichotomy in how these paradigms handle information integrity. RAG consistently excels in maintaining verifiable, source-grounded factual fidelity, making it indispensable for dynamic, knowledge-intensive tasks. Conversely, Fine-tuning is highly effective for stylistic adaptation and deeply embedding static domain nuances, but inherently risks catastrophic forgetting and increased hallucinations when exposed to novel facts. A key finding is that optimal data fidelity is rarely achieved through a single method, rather, state-of-the-art performance increasingly relies on hybrid architectures and agentic frameworks that synergise retrieval precision with parametric understanding.

To advance the development of trustworthy AI, future research must shift from isolated performance metrics to comprehensive evaluations of truthfulness, integrity, and representational fairness. We recommend prioritising the standardisation of data fidelity benchmarks and the development of robust, privacy-preserving augmentation techniques. Additionally, exploring continual learning algorithms that mitigate knowledge degradation will be crucial. Finally, the integration of human experts into active, autonomous retrieval loops-moving beyond passive evaluation-represents a promising frontier for ensuring that augmented LLMs can safely and reliably navigate complex, high-stakes environments.

Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
CL	Continual Learning
DP	Differential Privacy
EM	Exact Match
KG	Knowledge Graph
LLM	Large Language Model
LoRA	Low-Rank Adaptation
PEFT	Parameter-Efficient Fine-tuning
RAG	Retrieval-Augmented Generation
ReAct	Reasoning and Acting

References

1. Zhang, Z.; Han, X.; Liu, Z.; Jiang, X.; Sun, M.; Liu, Q. ERNIE: Enhanced Language Representation with Informative Entities. In Proceedings of the Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; Korhonen, A.; Traum, D.; Màrquez, L., Eds., Florence, Italy, 2019; pp. 1441–1451. https://doi.org/10.18653/v1/P19-1139.

 doi:10.20944/preprints202604.0717.v1

2. Sun, Y.; Wang, S.; Feng, S.; Ding, S.; Pang, C.; Shang, J.; Liu, J.; Chen, X.; Zhao, Y.; Lu, Y.; et al. ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation. *ArXiv* **2021**, *abs/2107.02137*.
3. Xue, J.; Zheng, M.; Hu, Y.; Liu, F.; Chen, X.; Lou, Q. BadRAG: Identifying Vulnerabilities in Retrieval Augmented Generation of Large Language Models. *CoRR* **2024**, *abs/2406.00083*, [2406.00083]. https://doi.org/10.48550/ARXIV.2406.00083.
4. Weihang, S.; Tang, Y.; Ai, Q.; Wu, Z.; Liu, Y. DRAGIN: Dynamic Retrieval Augmented Generation based on the Real-time Information Needs of Large Language Models. 01 2024, pp. 12991–13013. https://doi.org/10.18653/v1/2024.acl-long.702.
5. Chauhan, P.; Sahani, R.K.; Datta, S.; Qadir, A.; Raj, M.; Ali, M.M. Evaluating Top-k RAG-based approach for Game Review Generation. In Proceedings of the 2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT), 2024, Vol. 5, pp. 258–263. https://doi.org/10.1109/IC2PCT60090.2024.10486273.
6. Wang, C.; Long, Q.; Xiao, M.; Cai, X.; Wu, C.; Meng, Z.; Wang, X.; Zhou, Y. BioRAG: A RAG-LLM Framework for Biological Question Reasoning. *CoRR* **2024**, *abs/2408.01107*, [2408.01107]. https://doi.org/10.48550/ARXIV.2408.01107.
7. Salemi, A.; Zamani, H. Evaluating Retrieval Quality in Retrieval-Augmented Generation. In Proceedings of the Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, 2024; SIGIR '24, p. 2395–2400. https://doi.org/10.1145/3626772.3657957.
8. Spennemann, D.H.R. Generative Artificial Intelligence and the Future of Public Knowledge. *Knowledge* **2025**, *5*. https://doi.org/10.3390/knowledge5030020.
9. Koo, M. ChatGPT Research: A Bibliometric Analysis Based on the Web of Science from 2023 to June 2024. *Knowledge* **2025**, *5*. https://doi.org/10.3390/knowledge5010004.
10. Ovadia, O.; Brief, M.; Mishaeli, M.; Elisha, O. Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2023.
11. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Guo, Q.; Wang, M.; et al. Retrieval-Augmented Generation for Large Language Models: A Survey. *CoRR* **2023**, *abs/2312.10997*, [2312.10997]. https://doi.org/10.48550/ARXIV.2312.10997.
12. Azaria, A.; Mitchell, T. The Internal State of an LLM Knows When It's Lying, 2023, [arXiv:cs.CL/2304.13734].
13. Raj, G.; Hamzah.; Raj, N.; Ranjan, N. Hacking LLMs: A Technical Analysis of Security Vulnerabilities and Defense Mechanisms. In Proceedings of the 2025 2nd International Conference on Computational Intelligence, Communication Technology and Networking (CICTN), 2025, pp. 555–560. https://doi.org/10.1109/CICTN64563.2025.10932638.
14. Fasha, M.; Rub, F.A.; Matar, N.; Sowan, B.; Al Khaldy, M.; Barham, H. Mitigating the OWASP Top 10 For Large Language Models Applications using Intelligent Agents. In Proceedings of the 2024 2nd International Conference on Cyber Resilience (ICCR), 2024, pp. 1–9. https://doi.org/10.1109/ICCR61006.2024.10532874.
15. Gummadi, V.; Udayaraju, P.; Sarabu, V.R.; Ravulu, C.; Seelam, D.R.; Venkataramana, S. Enhancing Communication and Data Transmission Security in RAG Using Large Language Models. In Proceedings of the 2024 4th International Conference on Sustainable Expert Systems (ICSES), 2024, pp. 612–617. https://doi.org/10.1109/ICSES63445.2024.10763024.
16. M, L.; Agarwal, V.; Kamthania, S.; Vutkur, P.; S, M.C. Benchmarking LLM for Zero-day Vulnerabilities. In Proceedings of the 2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), 2024, pp. 1–6. https://doi.org/10.1109/CONECCT62155.2024.10677338.
17. Afsharmazayejani, R.; Shahmiri, M.M.; Link, P.; Pearce, H.; Tan, B. Toward Hardware Security Benchmarking of LLMs. In Proceedings of the 2024 IEEE LLM Aided Design Workshop (LAD), 2024, pp. 1–7. https://doi.org/10.1109/LAD62341.2024.10691745.
18. Rathod, V.; Nabavirazavi, S.; Zad, S.; Iyengar, S.S. Privacy and Security Challenges in Large Language Models. In Proceedings of the 2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC), 2025, pp. 00746–00752. https://doi.org/10.1109/CCWC62904.2025.10903912.
19. Pan, Q.; Wu, J. Selective Privacy-Preserving Federated Learning for Large Language Model Fine-Tuning. In Proceedings of the 2025 International Wireless Communications and Mobile Computing (IWCMC), 2025, pp. 1626–1631. https://doi.org/10.1109/IWCMC65282.2025.11059634.

 doi:10.20944/preprints202604.0717.v1

20. Gallegos, I.O.; Rossi, R.A.; Barrow, J.; Tanjim, M.M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; Ahmed, N.K. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics* **2024**, *50*, 1097–1179, [https://direct.mit.edu/coli/article-pdf/50/3/1097/2471010/coli_a_00524.pdf]. https://doi.org/10.1162/coli_a_00524.
21. Sharma, N.; Liao, Q.V.; Xiao, Z. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. In Proceedings of the Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, New York, NY, USA, 2024; CHI '24. https://doi.org/10.1145/3613904.3642459.
22. Sakurai, T.; Shiramatsu, S.; Kinoshita, R. LLM-based Agent for Recommending Information Related to Web Discussions at Appropriate Timing. In Proceedings of the 2024 IEEE International Conference on Agents (ICA), 2024, pp. 120–123. https://doi.org/10.1109/ICA63002.2024.00033.
23. Zhang, W.; Liu, H.; Dong, Z.; Du, Y.; Zhu, C.; Song, Y.; Zhu, H.; Wu, Z. Bridging the Information Gap Between Domain-Specific Model and General LLM for Personalized Recommendation. In Proceedings of the Web and Big Data; Zhang, W.; Tung, A.; Zheng, Z.; Yang, Z.; Wang, X.; Guo, H., Eds., Singapore, 2024; pp. 280–294.
24. Jin, B.; Yoon, J.; Han, J.; Arik, S.O. Long-Context LLMs Meet RAG: Overcoming Challenges for Long Inputs in RAG, 2024, [arXiv:cs.CL/2410.05983].
25. Jiang, Z.; Ma, X.; Chen, W. LongRAG: Enhancing Retrieval-Augmented Generation with Long-context LLMs, 2024, [arXiv:cs.CL/2406.15319].
26. Rivera, J.; Zapata, S.; Pizarro, R.; Keith, B. Enhancing Chatbot Performance in a SaaS Platform Through Retrieval-Augmented Generation and Prompt Engineering: A Case Study in Behavioral Safety Analysis. *Knowledge* **2025**, *5*. https://doi.org/10.3390/knowledge5040025.
27. Zhang, J.; Cui, W.; Huang, Y.; Das, K.; Kumar, S. Synthetic Knowledge Ingestion: Towards Knowledge Refinement and Injection for Enhancing Large Language Models. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; Al-Onaizan, Y.; Bansal, M.; Chen, Y.N., Eds., Miami, Florida, USA, 2024; pp. 21456–21473. https://doi.org/10.18653/v1/2024.emnlp-main.1196.
28. Yu, Y.; Ping, W.; Liu, Z.; Wang, B.; You, J.; Zhang, C.; Shoeybi, M.; Catanzaro, B. RankRAG: unifying context ranking with retrieval-augmented generation in LLMs. In Proceedings of the Proceedings of the 38th International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2025; NIPS '24.
29. Rackauckas, Z. Rag-Fusion: A New Take on Retrieval Augmented Generation. *International Journal on Natural Language Computing* **2024**, *13*, 37–47. https://doi.org/10.5121/ijnlc.2024.13103.
30. Hu, E.J.; yelong shen.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In Proceedings of the International Conference on Learning Representations, 2022.
31. Dong, H.; Xiong, W.; Goyal, D.; Pan, R.; Diao, S.; Zhang, J.; Shum, K.; Zhang, T. RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment. *ArXiv* **2023**, *abs/2304.06767*.
32. Yin, F.; Ye, X.; Durrett, G. LOFIT: localized fine-tuning on LLM representations. In Proceedings of the Proceedings of the 38th International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2025; NIPS '24.
33. Tonolini, F.; Massiah, J.; Aletras, N.; Kazai, G. Multi-Fidelity Fine-Tuning of Pre-Trained Language Models, 2024.
34. Pillai, S. Replay to Remember: Retaining Domain Knowledge in Streaming Language Models, 2025, [arXiv:cs.LG/2504.17780].
35. Hu, Y.; Lei, Z.; Zhang, Z.; Pan, B.; Ling, C.; Zhao, L. GRAG: Graph Retrieval-Augmented Generation. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2025; Chiruzzo, L.; Ritter, A.; Wang, L., Eds., Albuquerque, New Mexico, 2025; pp. 4145–4157. https://doi.org/10.18653/v1/2025.findings-naacl.232.
36. Peters, M.E.; Neumann, M.; Logan, R.; Schwartz, R.; Joshi, V.; Singh, S.; Smith, N.A. Knowledge Enhanced Contextual Word Representations. In Proceedings of the Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); Inui, K.; Jiang, J.; Ng, V.; Wan, X., Eds., Hong Kong, China, 2019; pp. 43–54. https://doi.org/10.18653/v1/D19-1005.

 doi:10.20944/preprints202604.0717.v1

37. Saeid, Y.; Kopinski, T. AgentFusion: A Multi-Agent Approach to Accurate Text Generation. In Proceedings of the 2024 International Conference on Electrical and Computer Engineering Researches (ICECER), 2024, pp. 1–8. https://doi.org/10.1109/ICECER62944.2024.10920460.
38. Parthasarathy, K.; Vaidhyanathan, K.; Dhar, R.; Krishnamachari, V.; Kakran, A.; Akshathala, S.; Arun, S.; Karan, A.; Muhammed, B.; Dubey, S.; et al. Engineering LLM Powered Multi-Agent Framework for Autonomous CloudOps. In Proceedings of the 2025 IEEE/ACM 4th International Conference on AI Engineering – Software Engineering for AI (CAIN), 2025, pp. 201–211. https://doi.org/10.1109/CAIN66642.2025.00031.
39. Afzal, A.; Vladika, J.; Fazlija, G.; Staradubets, A.; Matthes, F. Towards Optimizing a Retrieval Augmented Generation using Large Language Model on Academic Data. In Proceedings of the Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval, New York, NY, USA, 2025; NLPIR '24, p. 250–257. https://doi.org/10.1145/3711542.3711575.
40. Roychowdhury, S.; Krema, M.; Mahammad, A.; Moore, B.; Mukherjee, A.; Prakashchandra, P. ERATTA: Extreme RAG for enterprise-Table To Answers with Large Language Models. In Proceedings of the 2024 IEEE International Conference on Big Data (BigData), 2024, pp. 4605–4610. https://doi.org/10.1109/BigData62323.2024.10825910.
41. Gekhman, Z.; Yona, G.; Aharoni, R.; Eyal, M.; Feder, A.; Reichart, R.; Herzig, J. Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations? In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; Al-Onaizan, Y.; Bansal, M.; Chen, Y.N., Eds., Miami, Florida, USA, 2024; pp. 7765–7784. https://doi.org/10.18653/v1/2024.emnlp-main.444.
42. Jiang, X.; Fang, Y.; Qiu, R.; Zhang, H.; Xu, Y.; Chen, H.; Zhang, W.; Zhang, R.; Fang, Y.; Ma, X.; et al. TC–RAG: Turing–Complete RAG's Case study on Medical LLM Systems. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., Vienna, Austria, 2025; pp. 11400–11426. https://doi.org/10.18653/v1/2025.acl-long.558.
43. Ye, F.; Li, S.; Zhang, Y.; Chen, L. R2AG: Incorporating Retrieval Information into Retrieval Augmented Generation, 2024, [arXiv:cs.CL/2406.13249].
44. Du, X.; Zheng, G.; Wang, K.; Feng, J.; Deng, W.; Liu, M.; Chen, B.; Peng, X.; Ma, T.; Lou, Y. Vul-RAG: Enhancing LLM-based Vulnerability Detection via Knowledge-level RAG. *CoRR* **2024**, *abs/2406.11147*, [2406.11147]. https://doi.org/10.48550/ARXIV.2406.11147.
45. Huang, Y.; Sun, L.; Wang, H.; Wu, S.; Zhang, Q.; Li, Y.; Gao, C.; Huang, Y.; Lyu, W.; Zhang, Y.; et al. Position: TRUSTLLM: trustworthiness in large language models. In Proceedings of the Proceedings of the 41st International Conference on Machine Learning. JMLR.org, 2024, ICML'24.
46. Hagen, M.; Völske, M.; Göring, S.; Stein, B. Axiomatic Result Re-Ranking. In Proceedings of the Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, New York, NY, USA, 2016; CIKM '16, p. 721–730. https://doi.org/10.1145/2983323.2983704.
47. Dai, S.; Xu, C.; Xu, S.; Pang, L.; Dong, Z.; Xu, J. Bias and Unfairness in Information Retrieval Systems: New Challenges in the LLM Era. In Proceedings of the Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM, 2024, KDD '24, p. 6437–6447. https://doi.org/10.1145/3637528.3671458.
48. Razuvayevskaya, O.; Wu, B.; Leite, J.A.; Heppell, F.; Srba, I.; Scarton, C.; Bontcheva, K.; Song, X. Comparison between parameter-efficient techniques and full fine-tuning: A case study on multilingual news article classification. *PLOS ONE* **2024**, *19*, e0301738. https://doi.org/10.1371/journal.pone.0301738.
49. Wei, J.; Yang, C.; Song, X.; Lu, Y.; Hu, N.Z.; Huang, J.; Tran, D.; Peng, D.; Liu, R.; Huang, D.; et al. Long-form factuality in large language models. In Proceedings of the The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024.
50. Huang, C.W.; Chen, Y.N. FactAlign: Long-form Factuality Alignment of Large Language Models, 2024, [arXiv:cs.CL/2410.01691].
51. Madabushi, H.T. FS-RAG: A Frame Semantics Based Approach for Improved Factual Accuracy in Large Language Models, 2024, [arXiv:cs.CL/2406.16167].
52. Wang, Y.; Wang, M.; Manzoor, M.A.; Liu, F.; Georgiev, G.N.; Das, R.J.; Nakov, P. Factuality of Large Language Models: A Survey. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; Al-Onaizan, Y.; Bansal, M.; Chen, Y.N., Eds., Miami, Florida, USA, 2024; pp. 19519–19529. https://doi.org/10.18653/v1/2024.emnlp-main.1088.

 doi:10.20944/preprints202604.0717.v1

53. He, X.; Tian, Y.; Sun, Y.; Chawla, N.V.; Laurent, T.; LeCun, Y.; Bresson, X.; Hooi, B. G-retriever: retrieval-augmented generation for textual graph understanding and question answering. In Proceedings of the Proceedings of the 38th International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2025; NIPS '24.
54. Han, H.; Wang, Y.; Shomer, H.; Guo, K.; Ding, J.; Lei, Y.; Halappanavar, M.; Rossi, R.A.; Mukherjee, S.; Tang, X.; et al. Retrieval-Augmented Generation with Graphs (GraphRAG), 2025, [arXiv:cs.IR/2501.00309].
55. Gao, L.; Ma, X.; Lin, J.; Callan, J. Precise Zero-Shot Dense Retrieval without Relevance Labels. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Rogers, A.; Boyd-Graber, J.; Okazaki, N., Eds., Toronto, Canada, 2023; pp. 1762–1777. https://doi.org/10.18653/v1/2023.acl-long.99.
56. Tian, S.; Luo, Y.; Xu, T.; Yuan, C.; Jiang, H.; Wei, C.; Wang, X. KG-Adapter: Enabling Knowledge Graph Integration in Large Language Models through Parameter-Efficient Fine-Tuning. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024; Ku, L.W.; Martins, A.; Srikumar, V., Eds., Bangkok, Thailand, 2024; pp. 3813–3828. https://doi.org/10.18653/v1/2024.findings-acl.229.
57. Raza, S.; Raval, A.; Chatrath, V. MBIAS: Mitigating Bias in Large Language Models While Retaining Context. In Proceedings of the Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis; De Clercq, O.; Barriere, V.; Barnes, J.; Klinger, R.; Sedoc, J.; Tafreshi, S., Eds., Bangkok, Thailand, 2024; pp. 97–111. https://doi.org/10.18653/v1/2024.wassa-1.9.
58. Wang, Y.; Shi, X.; Zhao, X. MLLM4Rec: multimodal information enhancing LLM for sequential recommendation. *J. Intell. Inf. Syst.* **2025**, *63*, 745–761. https://doi.org/10.1007/s10844-024-00915-3.
59. Lee, C.; Roy, R.; Xu, M.; Raiman, J.; Shoeybi, M.; Catanzaro, B.; Ping, W. NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models, 2025, [arXiv:cs.CL/2405.17428].
60. Bin Islam, S.; Rahman, M.; Hossain, K.; Hoque, E.; Joty, S.; Parvez, M.R. Open-RAG: Enhanced Retrieval Augmented Reasoning with Open-Source Large Language Models. 01 2024, pp. 14231–14244. https://doi.org/10.18653/v1/2024.findings-emnlp.831.
61. Jiang, W.; Zhang, S.; Han, B.; Wang, J.; Wang, B.; Kraska, T. PipeRAG: Fast Retrieval-Augmented Generation via Adaptive Pipeline Parallelism. In Proceedings of the Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1, New York, NY, USA, 2025; KDD '25, p. 589–600. https://doi.org/10.1145/3690624.3709194.
62. Wang, L.; Chen, S.; Jiang, L.; Pan, S.; Cai, R.; Yang, S.; Yang, F. Parameter-efficient fine-tuning in large language models: a survey of methodologies. *Artif. Intell. Rev.* **2025**, *58*.
63. Seo, S.; Noh, S.; Lee, J.; Lim, S.; Lee, W.H.; Kang, H. REVECA: adaptive planning and trajectory-based validation in cooperative language agents using information relevance and relative proximity. In Proceedings of the Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence. AAAI Press, 2025, AAAI'25/IAAI'25/EAAI'25. https://doi.org/10.1609/aaai.v39i22.34496.
64. Chan, C.M.; Xu, C.; Yuan, R.; Luo, H.; Xue, W.; Guo, Y.; Fu, J. RQ-RAG: Learning to Refine Queries for Retrieval Augmented Generation. In Proceedings of the First Conference on Language Modeling, 2024.
65. Wei Jie, Y.; Ferdinan, T.; Kazienko, P.; Satapathy, R.; Cambria, E. Self-training Large Language Models through Knowledge Detection. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024; Al-Onaizan, Y.; Bansal, M.; Chen, Y.N., Eds., Miami, Florida, USA, 2024; pp. 15033–15045. https://doi.org/10.18653/v1/2024.findings-emnlp.883.
66. Cosme, D.; Galvao, A.; e Abreu, F.B. A Systematic Literature Review on LLM-Based Information Retrieval: The Issue of Contents Classification. In Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, 2024.
67. Choi, Y.; Asif, M.A.; Han, Z.; Willes, J.; Krishnan, R.G. Teaching LLMs How to Learn with Contextual Fine-Tuning, 2025, [arXiv:cs.LG/2503.09032].
68. Wu, Z.; Hao, Y.; Mou, L. ULPT: Prompt Tuning with Ultra-Low-Dimensional Optimization, 2025, [arXiv:cs.CL/2502.04501].
69. Ning, L.; Liu, L.; Wu, J.; Wu, N.; Berlowitz, D.; Prakash, S.; Green, B.; O'Banion, S.; Xie, J. User-LLM: Efficient LLM Contextualization with User Embeddings. In Proceedings of the Companion Proceedings of the ACM on Web Conference 2025, New York, NY, USA, 2025; WWW '25, p. 1219–1223. https://doi.org/10.1145/3701716.3715463.

 doi:10.20944/preprints202604.0717.v1

70. Zhang, Y.; Wu, Y.; Hua, W.; Lu, X.; Hu, X. Understanding Dynamic Diffusion Process of LLM-based Agents under Information Asymmetry, 2025, [arXiv:cs.MA/2502.13160].
71. Li, Y.; Tan, Z.; Xiao, W. LLM for Uniform Information Extraction Using Multi-task Learning Optimization. In Proceedings of the Web and Big Data. APWeb-WAIM 2024 International Workshops; Zhang, W.; Tung, A.; Zheng, Z.; Yang, Z.; Wang, X.; Guo, H., Eds., Singapore, 2025; pp. 17–29.
72. Cambon, A.; Hecht, B.; Edelman, B.; Ngwe, D.; Jaffe, S.; Heger, A.; Vorvoreanu, M.; Peng, S.; Hofman, J.; Farach, A.; et al. Early LLM-based Tools for Enterprise Information Workers Likely Provide Meaningful Boosts to Productivity. Technical Report MSR-TR-2023-43, Microsoft, 2023.
73. Purohit, S.; Chin, G.; Mackey, P.S.; Cottam, J.A. GraphAide: Advanced Graph-Assisted Query and Reasoning System. In Proceedings of the 2024 IEEE International Conference on Big Data (BigData), 2024, pp. 3485–3493. https://doi.org/10.1109/BigData62323.2024.10825705.
74. Patel, V.; Tejani, P.; Parekh, J.; Huang, K.; Tan, X. Developing A Chatbot: A Hybrid Approach Using Deep Learning and RAG. In Proceedings of the 2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2024, pp. 273–280. https://doi.org/10.1109/WI-IAT62293.2024.00043.
75. Renney, H.; Nethercott, M.; Williams, O.; Evetts, J.; Lang, J. Reimagining the Data Landscape: A Multi-Agent Paradigm for Data Interfacing. In Proceedings of the 2025 8th International Conference on Data Science and Machine Learning Applications (CDMA), 2025, pp. 114–119. https://doi.org/10.1109/CDMA61895.2025.00025.
76. Xu, J.; Wang, J.; Leung, J.; Gu, J. GRASP: Municipal Budget AI Chatbots for Enhancing Civic Engagement. In Proceedings of the 2024 IEEE International Conference on Big Data (BigData), 2024, pp. 7438–7442. https://doi.org/10.1109/BigData62323.2024.10825975.
77. Xu, X.; Zhang, D.; Liu, Q.; Lu, Q.; Zhu, L. Agentic RAG with Human-in-the-Retrieval. In Proceedings of the 2025 IEEE 22nd International Conference on Software Architecture Companion (ICSA-C), 2025, pp. 498–502. https://doi.org/10.1109/ICSA-C65153.2025.00074.
78. Honnalli, R.; Farooq, J. LLM-Powered Agentic AI Approach to Securing EV Charging Systems Against Cyber Threats. In Proceedings of the 2025 IEEE 26th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2025, pp. 266–274. https://doi.org/10.1109/WoWMoM65615.2025.00053.
79. Allan, K.; Azcona, J.; Sripada, S.; Leontidis, G.; Sutherland, C.A.M.; Phillips, L.H.; Martin, D. Stereotypical bias amplification and reversal in an experimental model of human interaction with generative artificial intelligence. *R. Soc. Open Sci.* **2025**, *12*, 241472.
80. Mondal, D.; Lipizzi, C. Mitigating Large Language Model Bias: Automated Dataset Augmentation and Prejudice Quantification. *Computers* **2024**, *13*. https://doi.org/10.3390/computers13060141.
81. Su, C.; Wen, J.; Kang, J.; Wang, Y.; Su, Y.; Pan, H.; Zhong, Z.; Shamim Hossain, M. Hybrid RAG-Empowered Multimodal LLM for Secure Data Management in Internet of Medical Things: A Diffusion-Based Contract Approach. *IEEE Internet of Things Journal* **2025**, *12*, 13428–13440. https://doi.org/10.1109/JIOT.2024.3521425.
82. Lei, Y.; Ding, L.; Cao, Y.; Zan, C.; Yates, A.; Tao, D. Unsupervised Dense Retrieval with Relevance-Aware Contrastive Pre-Training. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023; Rogers, A.; Boyd-Graber, J.; Okazaki, N., Eds., Toronto, Canada, 2023; pp. 10932–10940. https://doi.org/10.18653/v1/2023.findings-acl.695.
83. Shi, Y.; Zi, X.; Shi, Z.; Zhang, H.; Wu, Q.; Xu, M. Enhancing Retrieval and Managing Retrieval: A Four-Module Synergy for Improved Quality and Efficiency in RAG Systems. *ArXiv* **2024**, *abs/2407.10670*.
84. Yuan, X.J.; Guo, Q.; Dadson, Y.A.; Goodarzi, M.; Jung, J.; Dong, Y.; Albert, N.; Bennett Gayle, D.; Sharma, P.; Ogunbayo, O.T.; et al. A Review of Ethical Challenges in AI for Emergency Management. *Knowledge* **2025**, *5*. https://doi.org/10.3390/knowledge5030021.
85. Tsoulos, I.G.; Charilogis, V. Gen2Gen: Efficiently Training Artificial Neural Networks Using a Series of Genetic Algorithms. *Knowledge* **2025**, *5*. https://doi.org/10.3390/knowledge5030017.
86. Soudani, H.; Kanoulas, E.; Hasibi, F. Fine Tuning vs. Retrieval Augmented Generation for Less Popular Knowledge. In Proceedings of the Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, New York, NY, USA, 2024; SIGIR-AP 2024, p. 12–22. https://doi.org/10.1145/3673791.3698415.
87. Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.N.; Truitt, S.; Larson, J. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *ArXiv* **2024**, *abs/2404.16130*.

 doi:10.20944/preprints202604.0717.v1

88. Wang, S.; Xiang, Y. Research on Data Augmentation Techniques for Text Classification Based on Antonym Replacement and Random Swapping. In Proceedings of the Proceedings of the International Conference on Modeling, Natural Language Processing and Machine Learning, New York, NY, USA, 2024; CMNM '24, p. 103–108. https://doi.org/10.1145/3677779.3677796.
89. Liu, Y.; Cao, J.; Liu, C.; Ding, K.; Jin, L. Datasets for Large Language Models: A Comprehensive Survey. *ArXiv* **2024**, *abs/2402.18041*.
90. Mialon, G.; Dessì, R.; Lomeli, M.; Nalmpantis, C.; Pasunuru, R.; Raileanu, R.; Rozière, B.; Schick, T.; Dwivedi-Yu, J.; Celikyilmaz, A.; et al. Augmented Language Models: a Survey, 2023, [arXiv:cs.CL/2302.07842].
91. Nia, N.; Amiri, A.; Luo, Y.; Kline, A. Ethical perspectives on deployment of large language model agents in biomedicine: a survey. *AI and Ethics* **2025**, *6*. https://doi.org/10.1007/s43681-025-00847-w.