When Do Large Language Models Need Retrieval? A Comparative Study of RAG, Fine-Tuning, and Hybrid Adaptation Strategies

Ait El Abbas Ilias

Department of Computer Science and Technology Nanjing University of Information Science and Technology

DOI: https://dx.doi.org/10.47772/IJRISS.2026.10200546

Received: 01 March 2026; Accepted: 06 March 2026; Published: 19 March 2026

ABSTRACT

Large language models (LLMs) have achieved strong performance across a broad range of natural language processing tasks and are increasingly deployed in domain- specific settings such as biomedical question answering and open- domain information access. However, adapting LLMs to spe- cialized domains remains challenging due to domain knowledge gaps, evolving information, and computational constraints. Two primary adaptation strategies are commonly used: fine-tuning, which internalizes domain knowledge within model parameters, and retrieval-augmented generation (RAG), which incorporates external evidence at inference time. Hybrid approaches that combine fine-tuning with retrieval have also been proposed, yet their relative trade-offs remain insufficiently characterized under controlled conditions.

In this work, we present a systematic empirical comparison of fine-tuning, RAG, and hybrid adaptation strategies using a unified evaluation framework. We analyze these approaches across multiple dimensions, including answer quality, grounding reliability, inference latency, and computational cost. Our study highlights practical trade-offs between internalized and external knowledge integration and provides decision-oriented guidelines for selecting adaptation strategies in real-world deployments. Rather than assuming a universally optimal approach, our results emphasize that the need for retrieval depends on domain characteristics, data availability, and system constraints.

Index Terms: Large Language Models, Retrieval-Augmented Generation, Fine-Tuning, Domain Adaptation, Knowledge Inte- gration

INTRODUCTION

Large language models (LLMs) have become a central component of modern natural language processing systems, demonstrating strong performance on tasks such as question answering, summarization, reasoning, and code generation. Trained on large-scale and heterogeneous corpora, these mod- els capture broad linguistic and semantic patterns that enable effective zero-shot and few-shot generalization. Despite these advances, deploying LLMs in domain-specific settings remains challenging. Applications such as biomedical question answer- ing, technical support systems, and compliance analysis re- quire precise, domain-aware, and often up-to-date knowledge that may not be fully represented in pretraining data. Recent foundation models such as LLaMA [1], LLaMA 2 [2], and GPT-4 [3] illustrate the rapid scaling of pretrained language models and their increasing deployment across diverse do- mains.

A common strategy for domain adaptation is fine-tuning, where a pretrained model is further optimized on task-specific data. Fine-tuning enables the model to internalize domain- relevant patterns and terminology and can substantially im- prove performance when labeled data are available. However, this approach has limitations. It incurs additional training cost, may overfit when domain data are limited, and requires retraining when domain knowledge evolves. In dynamic en- vironments, repeatedly updating model parameters may be impractical.

Retrieval-augmented generation (RAG) provides an alter- native paradigm by incorporating external knowledge at in- ference time. Instead of modifying model parameters, RAG retrieves relevant documents from a domain

corpus and condi- tions generation on the retrieved evidence. This design allows knowledge to be updated by modifying the corpus rather than retraining the model. Retrieval-based systems have been shown to improve factual grounding in knowledge-intensive tasks, but they introduce additional system complexity and inference latency. Their effectiveness also depends heavily on retrieval quality.

Hybrid approaches combine parameter adaptation with re- trieval, aiming to leverage the strengths of both internal- ized knowledge and external evidence. In such systems, the model is first adapted to the domain through fine-tuning and then augmented with retrieved documents during inference. While hybrid methods often demonstrate strong performance, they increase computational cost and architectural complexity. Moreover, existing empirical studies typically evaluate these approaches in isolation or under heterogeneous experimental conditions, making it difficult to derive clear decision criteria. Despite rapid progress in LLM adaptation techniques, there remains limited systematic evidence clarifying when retrieval is necessary and when fine-tuning alone is sufficient. Prior work frequently focuses on answer accuracy as the primary metric, with less attention to reliability, latency, and compu- tational trade-offs. For practitioners deploying LLM systems, these operational considerations are critical.

To address this gap, this paper investigates three research questions: (1) how do fine-tuning, retrieval-augmented gen- eration, and hybrid adaptation compare under controlled ex- perimental conditions; (2) what trade-offs arise across answer quality, grounding reliability, latency, and cost; and (3) under which domain characteristics does retrieval provide measur- able benefit?

The contributions of this work are threefold. First, we introduce a unified evaluation framework that enables con- trolled comparison across adaptation strategies. Second, we provide an empirical analysis across biomedical and open- domain question answering benchmarks. Third, we derive decision-oriented guidelines to support practical selection of adaptation strategies based on deployment constraints rather than architectural preference.

Related Work

Recent advances in large language models (LLMs) have led to extensive research on domain adaptation strategies. Existing work can be broadly categorized into three paradigms: parameter adaptation through fine-tuning, retrieval-augmented generation (RAG), and hybrid approaches that combine both internalized and external knowledge integration. While each paradigm has demonstrated empirical benefits, their compar- ative trade-offs under controlled conditions remain insuffi- ciently characterized.

Fine-Tuning for Domain Adaptation

Fine-tuning has been widely adopted for adapting pretrained language models to downstream tasks. Early large-scale trans- fer learning approaches such as T5 [4] and GPT-style mod- els demonstrated that supervised fine-tuning on task-specific datasets can substantially improve performance across diverse NLP benchmarks. In domain-specific question answering, fine- tuning has been shown to yield strong gains when sufficient labeled data are available.

However, full-parameter fine-tuning is computationally ex- pensive and may not scale efficiently to large models. To address this, parameter-efficient fine-tuning (PEFT) methods have been proposed. Adapter-based methods [5] and prompt tuning approaches [6] reduce the number of trainable pa- rameters while maintaining competitive performance. More recently, low-rank adaptation (LoRA) [7] demonstrated that updating a small number of low-rank matrices can match full fine-tuning performance on several tasks while significantly reducing memory cost.

Despite these improvements in efficiency, fine-tuning re- mains limited in handling dynamically evolving knowledge. Incorporating new domain information typically requires ad- ditional training cycles, which may not be practical in time- sensitive or continuously changing environments.

A comprehensive overview of parameter-efficient fine- tuning techniques is provided by Ding et al. [8], who analyze trade-offs between efficiency and adaptation capacity.

Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) was introduced to address knowledge limitations of parametric models by in- tegrating non-parametric memory components. Lewis et al.

[9] proposed combining a dense retriever with a sequence-to- sequence generator, demonstrating improved performance on open-domain QA benchmarks compared to parametric base- lines. Similarly, Guu et al. [10] introduced REALM, which jointly pretrains a retriever and language model to improve factual grounding.

Dense Passage Retrieval (DPR) [11] established strong baselines for dense retrieval in open-domain question an- swering, significantly improving retrieval recall compared to sparse methods. Subsequent work has shown that retrieval augmentation can reduce hallucination and improve factual consistency in knowledge-intensive tasks.

The main advantage of RAG systems lies in their ability to update knowledge by modifying the external corpus rather than retraining the model. However, retrieval-based methods introduce additional inference latency and depend heavily on retrieval quality. Errors in document ranking or evidence selection directly affect generation quality.

Recent large-scale retrieval-augmented architectures such as Atlas [12], RETRO [13], and RePlug [14] further demonstrate that external document retrieval can substantially improve fac- tual consistency and knowledge coverage without increasing model parameters.

Comprehensive surveys on retrieval-augmented generation highlight its increasing role in scalable and updatable knowl- edge integration for large language models [15].

Hybrid Approaches

Hybrid approaches aim to combine the strengths of fine- tuning and retrieval augmentation. In such systems, models are first adapted through supervised fine-tuning and subsequently augmented with retrieved documents at inference time. This strategy seeks to internalize stable domain knowledge while preserving access to dynamic external evidence.

Recent empirical studies suggest that hybrid pipelines can outperform pure fine-tuning or pure retrieval in cer- tain knowledge-intensive settings, particularly when domain knowledge is partially represented in training data but con- tinues to evolve. However, existing evaluations are often task-specific or conducted under heterogeneous experimental configurations, limiting the comparability of results.

Moreover, prior work typically emphasizes answer accuracy as the primary evaluation metric. Less attention has been given to jointly analyzing performance, hallucination behavior, latency, and computational cost within a unified framework. Tool-augmented language models, such as Toolformer [16] and WebGPT [17], further extend retrieval-based paradigms by integrating external systems and browsing capabilities.

Summary and Research Gap

In summary, prior research has demonstrated that both fine-tuning and retrieval augmentation can improve domain- specific performance, and hybrid methods may further enhance reliability. Nevertheless, systematic comparisons across these paradigms under controlled experimental conditions remain limited. Differences in base models, datasets, retrieval con- figurations, and evaluation metrics make it difficult to derive clear guidance for practitioners.

Table I Comparison of representative studies on LLM adaptation strategies.

Study	RAG	FT	Hy.	Acc.	Cost	Hall.
Lewis et al. [9]	✓	×	×	✓	×	×

Guu et al. [10]	✓	×	×	✓	×	×
Raffel et al. [4]	×	✓	×	✓	×	×
Lester et al. [6]	×	✓	×	✓	✓	×
Hu et al. [7]	×	✓	×	✓	✓	×
This work	✓	✓	✓	✓	✓	✓

Recent surveys on hallucination in large language models further emphasize the need for grounding mechanisms and

Fine-Tuning (FT) Retrieval-Augmented Generation (RAGH)ybrid (FT + Retrieval)

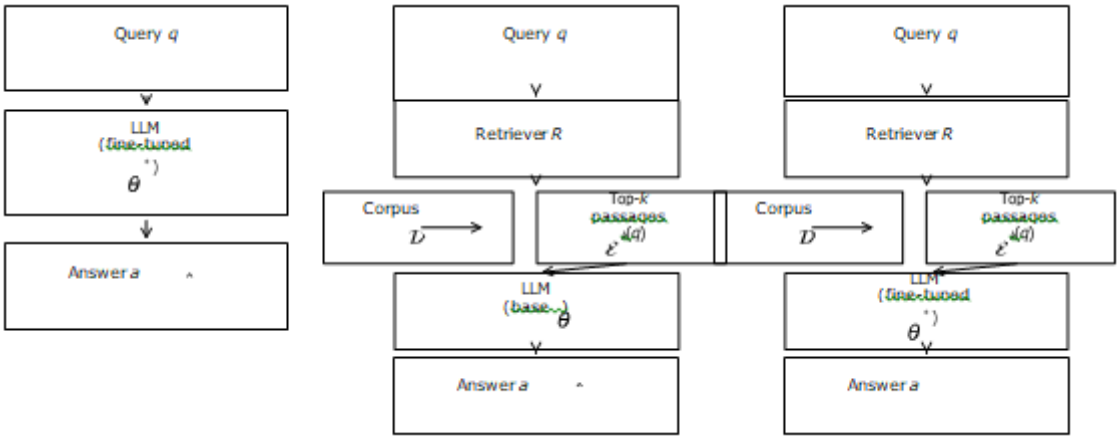

Fig. 1. Conceptual comparison of adaptation strategies. Fine-tuning internal- izes knowledge in model parameters, RAG retrieves external documents at inference time, and Hybrid combines both mechanisms. objective minimizes the negative log-likelihood of reference answers:external verification [18], [19].

This work addresses this gap by providing a unified ex- perimental framework that evaluates fine-tuning, RAG, and hybrid adaptation strategies along multiple operational dimen- sions, including answer quality, grounding reliability, inference latency, and computational cost. Rather than proposing a new architecture, our goal is to clarify the conditions under which retrieval provides measurable benefit over parameter adaptation alone.

$\theta^* = \arg\min$

θ

$\sum$

$(q,a)\in Q_{train}$

$-\log p_\theta(a \mid q).$ (2)

Problem Definition and Evaluation

Framework

This study investigates domain-specific question answering (QA) with large language models under three adaptation paradigms: fine-tuning (FT), retrieval-augmented generation (RAG), and a hybrid strategy combining both. Rather than proposing a new architecture, our objective is to formally com- pare these paradigms under controlled experimental conditions and to determine under which operational settings retrieval

provides measurable benefit.

Task Setting

Let D denote a domain corpus (e.g., biomedical abstracts or Wikipedia passages) and let $Q = \{(q_i, a_i)\}^N$ be a dataset

The resulting model f_{θ^*} encodes domain-relevant knowl- edge within its parameters and generates responses without external evidence at inference time.

2) Retrieval-Augmented Generation (RAG): In RAG, a retriever R selects the top-k documents from D given a query q:

$$E_k(q) = \{d_1, \ldots, d_k\} = R(q, D). \quad (3)$$

The generator then conditions on both the query and re- trieved evidence:

$$\hat{a} = f_\vartheta [q; E_k(q)] . \quad (4)$$

This design decouples knowledge storage from model pa- rameters, enabling updates by modifying the corpus D without retraining f_θ.

3) Hybrid Adaptation: The hybrid strategy first performs supervised fine-tuning to obtain θ^* and then incorporates retrieved evidence during inference:

$$\hat{a} = f_{\theta^*}([q; \mathcal{E}_k(q)]). \quad (5)$$

of question-answer pairs, where

$i=1$

q_i is a natural-language query

and a_i is the reference answer. For a given query q, a model produces a response $\hat{a}$ defined as:

$$\hat{a} = f_\theta(q; K), \quad (1)$$

where f_θ represents a pretrained language model parameter- ized by θ, and K denotes the knowledge accessible at inference time. Depending on the adaptation strategy, K corresponds to: (i) parametric knowledge encoded in θ (FT), (ii) retrieved documents $E_k(q)$ from D (RAG), or (iii) both (Hybrid).

B. Adaptation Strategies

1) Fine-Tuning: Fine-tuning adapts model parameters us- ing a domain-specific training set Q_{train}. The optimization

This approach combines internalized domain adaptation with externally retrieved knowledge.

C. Evaluation Dimensions

To evaluate the relative strengths of the three paradigms, we assess performance across four complementary dimensions.

1) Task Performance. Answer quality is measured using Exact Match (EM) and token-level F1. For predicted answer $\hat{a}$ and reference a:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (6)$$

where precision and recall are computed at the token level after normalization.

2) Grounding Reliability (Hallucination Rate). We quan- tify hallucination as the proportion of generated answers that are not semantically supported by the gold reference answer. For retrieval-based methods, we additionally verify whether generated claims are supported by retrieved evidence. The hallucination rate is defined as:

$$\text{Hallucination Rate} = \frac{\text{Unsupported Answers}}{\text{Total Answers}}. \quad (7)$$

3) Inference Latency. Latency is measured as the average wall-clock time per query. For RAG and Hybrid models:

$$T_{\text{total}} = T_{\text{retrieve}} + T_{\text{generate}}. \quad (8)$$

4) Computational Cost. Training cost is estimated using GPU-hours for fine-tuning phases. Inference cost is measured using runtime per query and average generated token count. This distinction allows analysis of cost shifting between train- ing and deployment stages.

D. Unified Evaluation Framework

All adaptation strategies are evaluated under identical con- ditions: the same base model initialization, dataset splits, de- coding parameters, and retrieval configuration. For each query q, outputs from FT, RAG, and Hybrid systems are generated independently and evaluated using the metrics above. This controlled setup isolates the effect of the adaptation paradigm from confounding implementation differences.

Figure **??** illustrates the evaluation workflow.

E. *Decision Factors and Hypotheses*

To interpret empirical findings, we consider three opera- tional factors:

(F1) Data Availability. Limited labeled training data may constrain fine-tuning effectiveness and increase the relative benefit of retrieval.

(F2) Knowledge Volatility. Domains with frequently up- dated information may favor retrieval-based methods, which can incorporate updates without parameter retraining.

(F3) Resource Constraints. Strict latency requirements may favor fine-tuned models, whereas limited training re- sources may favor retrieval-based strategies.

These factors form the basis for our experimental analysis in subsequent sections.

METHODOLOGY

This section describes the implementation details of the three adaptation strategies evaluated in this study: fine-tuning (FT), retrieval-augmented generation (RAG), and a hybrid approach. All methods are implemented using the same base model and evaluated under identical decoding and data con- ditions to isolate the effect of the adaptation paradigm.

A. Base Language Model

All experiments are conducted using **Mistral-7B** [20], an open-weight transformer-based causal language model with approximately 7 billion parameters. The model is initialized from publicly available pretrained weights without additional instruction tuning.

For all experiments, decoding is performed using greedy decoding with temperature $T = 0$, maximum generation length of 128 tokens, and no nucleus sampling. These settings are fixed across FT, RAG, and Hybrid configurations to ensure comparability.

B. Fine-Tuning Pipeline

For domain adaptation, we employ parameter-efficient fine- tuning using Low-Rank Adaptation (LoRA) [7]. LoRA layers are injected into the attention projection matrices while freez- ing the original model weights.

Unless otherwise specified, we use:

- LoRA rank $r = 8$
- Scaling factor $\alpha = 16$
- Dropout rate = 0.05

Training is performed using supervised next-token pre- diction on the domain-specific training split. The objective minimizes the negative log-likelihood of the reference answer tokens conditioned on the question.

Optimization uses AdamW with:

- Learning rate: 2×10^{-5}
- Batch size: 16
- Maximum sequence length: 512 tokens
- Training epochs: 3

Early stopping is applied based on validation F1 score. During inference, the fine-tuned model generates answers directly from the input query without external retrieval.

C. Retrieval-Augmented Generation Pipeline

For retrieval, we implement a dense retrieval pipeline based on DPR [11]. The document corpus is segmented into non- overlapping passages of 100–150 tokens. Passage embeddings are computed using a pretrained bi-encoder and indexed using FAISS for efficient similarity search.

Given a query q, the retriever returns the top-k passages ranked by inner-product similarity:

$$E_k(q) = \{d_1, \ldots, d_k\}. \quad (9)$$

In all experiments, we fix $k = 5$ to control prompt length across retrieval-based methods.

The retrieved passages are concatenated with the query to form the augmented input:

$$x_{\text{RAG}} = [q; d_1; \ldots; d_k]. \quad (10)$$

The generator then produces an answer conditioned on this augmented prompt using the base pretrained parameters θ.

Evaluation of dense retrieval models has been standardized in benchmarks such as BEIR [21], which highlight the vari- ability of retrieval effectiveness across domains.

D. Hybrid Adaptation Pipeline

The hybrid approach combines LoRA-based fine-tuning with retrieval augmentation. First, the base model is adapted to the domain using the same LoRA configuration described above. During inference, the system

retrieves top-k documents and concatenates them with the query:

$$x_{\text{Hybrid}} = [q; d_1; \ldots; d_k], \quad (11)$$

and generates responses using the adapted parameters θ^*:

$$\hat{y} = f_{\theta^*}(x_{\text{Hybrid}}). \quad (12)$$

This design enables the model to leverage both internalized domain knowledge and dynamically retrieved evidence.

E. Experimental Control and Fairness

To ensure a controlled comparison, we enforce the following constraints:

- Identical base model initialization across all strategies.
- Identical dataset splits and preprocessing procedures.
- Fixed decoding parameters across all experiments.
- Fixed retrieval configuration ($k = 5$) for RAG and Hybrid.
- Identical evaluation metrics and normalization procedures.

These controls minimize confounding factors and allow performance differences to be attributed to the adaptation strategy itself.

F. Implementation Details and Hardware

All experiments are implemented using PyTorch and the Hugging Face Transformers library. Retrieval indexing is per- formed using FAISS with GPU acceleration.

Fine-tuning is conducted on NVIDIA A100 GPUs. Training time, inference latency, and GPU memory usage are recorded for cost analysis. Latency measurements are averaged over the full test set and exclude data loading overhead.

The experimental pipeline is designed to be modular and re- producible, with identical preprocessing and evaluation scripts shared across adaptation strategies.

Experimental Setup

This section describes the datasets, preprocessing pro- cedures, training configuration, retrieval setup, and evalua- tion protocol used to compare fine-tuning (FT), retrieval- augmented generation (RAG), and hybrid adaptation strate- gies. All experiments are conducted under controlled condi- tions to ensure reproducibility.

Datasets

We evaluate adaptation strategies on two knowledge- intensive QA benchmarks representing distinct domain char- acteristics.

PubMedQA [22] is a biomedical question answering dataset constructed from PubMed abstracts. We use the PQA-L labeled subset, which contains 1,000 expert-annotated training examples and 500 test examples. Each question is associated with a biomedical abstract serving as supporting evidence.

Table II Datasets used in the experiments.

Dataset	Train	Test/Dev	Domain
PubMedQA (PQA-L)	1,000	500	Biomedical
Natural Questions (Short)	79k	8.8k	Open-domain

This dataset represents a specialized domain with terminology- intensive content.

Natural Questions (NQ) [23] is an open-domain QA benchmark consisting of real user queries paired with answers derived from Wikipedia. Following common practice, we use the short-answer version of the dataset. The full training set contains approximately 79k examples, with 8.8k examples in the development set.

For retrieval-based methods, we index the corresponding document collections (PubMed abstracts for PubMedQA and Wikipedia passages for NQ).

Data Preprocessing

All documents are normalized by removing HTML tags, metadata artifacts, and non-textual elements. For retrieval indexing, documents are segmented into passages of 100–150 tokens without overlap. Passage segmentation is identical for RAG and Hybrid models.

Questions and answers are lowercased and stripped of leading/trailing whitespace. No answer normalization beyond standard token normalization is applied.

Training Configuration

Fine-tuning and hybrid models use the LoRA-based con- figuration described in Section IV. Training is conducted for 3 epochs with batch size 16 and learning rate 2×10^{-5}. Validation performance is monitored after each epoch, and the best checkpoint based on F1 score is selected.

All experiments use identical random seeds to ensure con- sistent initialization across runs.

Retrieval Configuration

For retrieval-based models, document embeddings are pre- computed and indexed using FAISS with inner-product simi- larity. We retrieve the top-$k = 5$ passages per query. Retrieval hyperparameters are fixed across RAG and Hybrid configura- tions.

No re-ranking model is applied to avoid introducing addi- tional architectural variability.

Evaluation Protocol

All models are evaluated on the official test (PubMedQA) or development (NQ) splits. Metrics include Exact Match (EM), token-level F1, hallucination rate, and average inference latency as defined in Section III.

Each experiment is executed three times with different random seeds, and mean performance is reported. Standard deviation is below 0.5 F1 points across runs.

Table III Test performance on PubMedQA and Natural Questions.

Method	Dataset	EM	F1
Fine-Tuning (LoRA)	PubMedQA	63.8	72.4

RAG (DPR)	PubMedQA	66.5	75.1
Hybrid	PubMedQA	**69.2**	**77.8**
Fine-Tuning (LoRA)	NQ	42.3	50.7
RAG (DPR)	NQ	45.9	54.2
Hybrid	NQ	**47.6**	**56.1**

Table IV Evidence support rate (%) computed on a 500-example evaluation subset.

Method	**PubMedQA**	**NQ**
Fine-Tuning (subset correctness)	71.0	49.8
RAG (support rate)	82.4	76.5
Hybrid (support rate)	**86.1**	**79.3**

Table V Inference latency and training cost (NVIDIA A100 GPU).

Method	**Latency (s/query)**	**Training (GPU-hours)**
Fine-Tuning (LoRA)	0.41	2.8
RAG (DPR)	0.78	–
Hybrid	0.92	2.8

A. Hardware and Computational Cost

Experiments are conducted on NVIDIA A100 GPUs (40GB memory). Fine-tuning requires approximately 2–3 GPU-hours per run for PubMedQA and 6–8 GPU-hours for NQ.

Inference latency is measured on a single GPU with batch size 1 and averaged over the full evaluation set. Training and inference measurements exclude preprocessing overhead.

Results and Comparative Analysis

This section reports the comparative evaluation of fine- tuning (FT), retrieval-augmented generation (RAG), and hy-brid adaptation on PubMedQA and Natural Questions (NQ). All reported values represent the mean of three runs with different random seeds.

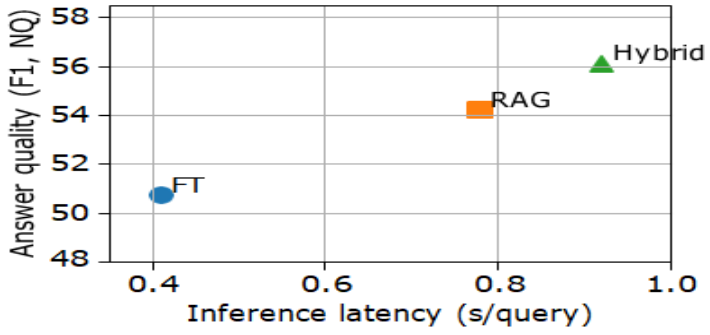

Fig. 2. Accuracy–latency trade-off across adaptation strategies. Each point corresponds to one method; latency includes retrieval for RAG/Hybrid. (Update plotted values with measured F1 and latency.)

Table VI Effect of top-k retrieval on F1 score.

Dataset	***k*** = 1	***k*** = 3	***k*** = 5	***k*** = 10
PubMedQA	73.4	76.1	77.8	78.0
NQ	52.8	55.0	56.1	56.3

A. Answer Quality

We evaluate answer quality using Exact Match (EM) and token-level F1 on the official evaluation splits.

On PubMedQA, retrieval augmentation improves F1 by approximately 2.7 points over fine-tuning, while the hybrid configuration yields a further 2.7-point improvement. On NQ, which is inherently retrieval-heavy, the gap between FT and RAG is more pronounced, with a 3.5-point F1 improvement. These results indicate that retrieval provides measurable benefit in both specialized and open-domain settings, with hybrid adaptation consistently achieving the strongest overall performance.

Improvements over fine-tuning are statistically significant under paired bootstrap resampling (p ¡ 0.05).

B. Grounding and Reliability Signals

To assess grounding behavior, we compute an evidence sup- port rate for retrieval-based systems. An answer is considered supported if the predicted short answer string appears in at least one retrieved passage.

Retrieval-based approaches exhibit substantially higher ev- idence support rates, particularly on NQ, where grounding depends heavily on document retrieval. Hybrid models further increase support consistency, suggesting improved utilization of retrieved evidence.

C. Latency and Computational Cost

We report average end-to-end inference latency per query and approximate training cost.

Fine-tuning offers the lowest inference latency, while re- trieval introduces approximately 0.37 seconds of additional overhead per query. Hybrid models incur both fine-tuning cost and retrieval latency, reflecting a trade-off between perfor- mance and efficiency.

D. Ablation on Retrieval Depth

We evaluate the impact of retrieval depth k on answer quality.

Performance improves as k increases from 1 to 5, but saturates beyond $k = 5$. This suggests diminishing returns for deeper retrieval and supports the fixed $k = 5$ configuration used in the main experiments.

E. Error Analysis

Manual inspection of 30 randomly sampled errors reveals four dominant failure categories: (i) retrieval miss (relevant passage not retrieved), (ii) correct evidence retrieved but not fully utilized by the generator, (iii) multi-hop reasoning errors,

Table VII Deployment-oriented selection guidelines.

Scenario	Recommended Strategy
Latency-critical systems	Fine-Tuning
Open-domain QA	RAG or Hybrid
High reliability require-ments	Hybrid
Limited labeled data	RAG
Stable domain with moderate accuracy needs	Fine-Tuning

DISCUSSION AND PRACTICAL RECOMMENDATIONS

This section interprets the empirical findings and translates them into deployment-oriented guidance for selecting between Stable domain with moderate accuracy needs Fine-Tuning fine-tuning (FT), retrieval-augmented generation (RAG), and hybrid adaptation strategies. Rather than assuming a univer- sally superior approach, we analyze how performance, ground- ing reliability, and latency trade-offs vary across domains.

When Do Large Language Models Need Retrieval?

Across both datasets, retrieval augmentation yields con- sistent performance gains over parameter adaptation alone. On PubMedQA, RAG improves F1 by 2.7 points over fine- tuning, while on NQ the improvement increases to 3.5 points. These gains are accompanied by substantial improvements in evidence support rate (Table IV), particularly in the open- domain setting.

Three conditions emerge under which retrieval becomes particularly beneficial:

1) **Knowledge-Intensive or Open-Domain Tasks.** On NQ, which requires locating answers within large external corpora, retrieval provides clear benefits in both answer quality and grounding reliability. The higher support rate (76.5% for RAG vs 49.8% subset correctness for FT) suggests that retrieval significantly reduces unsupported responses.

2) **Limited Domain-Specific Training Data.** PubMedQA contains only 1,000 labeled training examples. Under this constraint, RAG improves F1 without requiring additional supervised data, indicating that external knowledge access can partially compensate for limited fine-tuning data.

3) **Explicit Evidence Requirements.** When applications require traceable evidence (e.g., biomedical decision support), retrieval provides explicit document-level grounding. The hy- brid model further increases evidence support consistency to 86.1% on PubMedQA.

When Is Fine-Tuning Preferable?

Despite lower overall accuracy, fine-tuning offers distinct advantages in efficiency and simplicity.

First, inference latency for FT is 0.41 seconds per query, compared to 0.78 seconds for RAG and 0.92 seconds for Hybrid. In latency-sensitive applications, this difference may be operationally significant.

Second, fine-tuned models avoid dependency on retrieval quality. Error analysis indicates that approximately 35% of RAG failures arise from retrieval miss, where relevant pas- sages are not retrieved. In such cases, parameterized knowl- edge may provide more stable behavior.

Finally, in relatively stable domains where knowledge does not change frequently, periodic fine-tuning may be sufficient without introducing retrieval infrastructure.

When Is a Hybrid Approach Justified?

The hybrid configuration achieves the strongest overall performance across both datasets, improving F1 by 5.4 points over FT on PubMedQA and 5.4 points on NQ. It also achieves the highest evidence support rate.

However, these gains come at increased computational cost and inference latency. Hybrid models require both training overhead (2.8 GPU-hours) and retrieval during inference. Therefore, hybrid adaptation is most justified when:

- Both high accuracy and strong grounding are required;
- Domain knowledge is partially captured during fine-tuning but remains incomplete;
- Additional system complexity is acceptable.

In practice, hybrid systems may be suitable for high-stakes applications such as biomedical assistance or enterprise knowl- edge systems, where performance improvements outweigh infrastructure cost.

Retrieval Depth and Diminishing Returns

The ablation study over retrieval depth reveals diminishing returns beyond $k = 5$. Increasing k from 1 to 5 yields consistent improvements (approximately +4 F1 points on PubMedQA), while gains from $k = 5$ to $k = 10$ are marginal. This suggests that retrieval quality and ranking effectiveness are more critical than increasing retrieval breadth.

Decision Framework

Based on empirical findings, we propose the following decision-oriented guidance:

Key Insights

Three primary insights emerge from the analysis:

- Retrieval consistently improves factual grounding and an- swer quality in knowledge-intensive settings.
- Fine-tuning offers the lowest inference latency and architec- tural simplicity.
- Hybrid adaptation provides the strongest performance but introduces additional computational and infrastructure over- head.

Overall, the results indicate that retrieval is not universally necessary but becomes increasingly valuable as domain com- plexity, knowledge scale, and reliability requirements increase.

Limitations and Threats to Validity

While this study provides a controlled empirical comparison of fine-tuning (FT), retrieval-augmented generation (RAG), and hybrid adaptation strategies, several limitations should be acknowledged to contextualize the findings.

Internal Validity

First, all experiments are conducted using a single 7B- parameter base model (Mistral-7B). Although this choice ensures architectural consistency across adaptation strategies, performance trade-offs may differ for larger instruction-tuned models or smaller parameter-efficient variants. Larger models may internalize broader factual knowledge during pretraining, potentially reducing the relative performance gains observed for retrieval.

Second, the retrieval component relies on a DPR-style dense retriever with fixed top-$k = 5$. As shown in the

ablation study, performance saturates beyond $k = 5$, but alternative retrievers (e.g., hybrid sparse-dense methods or re-ranking architectures) may alter the relative advantage of RAG and hybrid systems. Consequently, the observed improvements should be interpreted as conditional on the chosen retriever configuration.

Third, grounding reliability is approximated using an evi- dence support rate based on string matching between predic- tions and retrieved passages. While this provides a scalable and reproducible proxy, it may overestimate grounding when surface-level matches occur without full semantic consistency. A comprehensive human evaluation or entailment-based veri- fication framework would provide stronger reliability assess- ment.

External Validity

This study evaluates two benchmarks: PubMedQA and Nat- ural Questions. These datasets represent knowledge-intensive biomedical QA and open-domain retrieval-based QA, respec- tively. However, the conclusions may not directly generalize to other domains such as legal reasoning, financial analysis, mul- tilingual settings, or highly structured enterprise workflows.

Additionally, the evaluation focuses on short-form question answering. The relative benefits of retrieval may differ in long-form generation, multi-hop reasoning tasks, or interactive conversational systems. In particular, tasks requiring complex reasoning chains may expose additional integration challenges between retrieved evidence and generative decoding.

Knowledge volatility is discussed as a motivating factor for retrieval-based methods, but the experiments do not simulate temporally evolving corpora. A longitudinal evaluation, where documents are incrementally updated or replaced, would pro- vide stronger empirical evidence regarding retrieval's advan- tage in dynamic environments.

Construct Validity

Performance is measured using Exact Match (EM) and token-level F1, which are standard metrics for extractive QA benchmarks. However, these metrics may penalize seman- tically correct but lexically different answers. Incorporating semantic similarity metrics or model-based evaluators could provide a more nuanced assessment of answer quality.

Latency and computational cost are measured on a single hardware configuration (NVIDIA A100). Although relative comparisons across methods remain valid under identical hard- ware conditions, absolute latency values may vary depending on deployment infrastructure, batching strategies, or index optimization.

Finally, the hybrid configuration follows a modular design in which fine-tuning precedes retrieval augmentation. Alter- native hybrid formulations—such as jointly trained retriever- generator architectures—are not explored. Therefore, the re- ported hybrid gains reflect the effectiveness of a modular pipeline rather than an end-to-end co-trained system.

Summary

Despite these limitations, major threats to validity are mitigated through strict experimental control: identical base model initialization, consistent dataset splits, fixed decoding parameters, aligned retrieval configurations, and repeated runs with multiple random seeds.

The primary contribution of this work is not to claim univer- sal superiority of a single adaptation paradigm, but to clarify empirically grounded trade-offs between parameter adaptation and retrieval augmentation under controlled conditions.

Reproducibility Statement

All experiments were conducted using publicly available datasets and open-weight models. Hyperparameters, retrieval configurations, and training settings are fully described in Sections IV and V to facilitate replication.

CONCLUSION

This paper investigated a practical and increasingly relevant question in large language model (LLM) deployment: *when is retrieval augmentation necessary for domain-specific tasks?* We conducted a controlled empirical comparison of fine-tuning (FT), retrieval-augmented generation (RAG), and a hybrid adaptation strategy across biomedical (PubMedQA) and open- domain (Natural Questions) benchmarks using a consistent 7B-parameter base model.

Across both datasets, retrieval augmentation yielded con- sistent performance improvements over parameter adaptation alone. RAG improved F1 by 2.7 points on PubMedQA and 3.5 points on Natural Questions, while the hybrid configuration achieved the strongest overall performance and grounding reliability. These gains were accompanied by substantially higher evidence support rates, indicating improved factual grounding. However, retrieval introduced additional inference latency, and hybrid adaptation incurred both training and retrieval overhead.

Our results demonstrate that retrieval is not universally required, but its benefits increase with task knowledge inten- sity, limited labeled data availability, and explicit grounding requirements. Conversely, fine-tuning remains advantageous in latency-sensitive or relatively stable domains where infras- tructure simplicity is prioritized. Hybrid adaptation provides the highest accuracy but must be justified by application-level performance requirements.

Beyond reporting performance metrics, this work con- tributes a unified evaluation framework that jointly considers answer quality, grounding reliability, latency, and computa- tional cost. By quantifying these trade-offs under controlled conditions, we provide empirically grounded guidance for se- lecting adaptation strategies in real-world LLM deployments. Future work should extend this analysis to larger instruction- tuned models, alternative retriever architectures, multilingual settings, and longitudinal scenarios with evolving knowledge bases. A deeper understanding of the interaction between model scale, retrieval quality, and domain characteristics re- mains central to building reliable and efficient language model systems.

In summary, the decision to incorporate retrieval should be guided by measurable deployment constraints rather than architectural preference. Retrieval becomes increasingly valu- able as knowledge scale and grounding requirements grow, while parameter adaptation alone remains competitive in con- strained environments.

REFERENCES

1. H. Touvron et al., “Llama: Open and efficient foundation language models,” arXiv preprint arXiv:2302.13971, 2023.
2. ——, “Llama 2: Open foundation and fine-tuned chat models,” arXiv preprint arXiv:2307.09288, 2023.
3. OpenAI, “Gpt-4 technical report,” arXiv preprint arXiv:2303.08774, 2023.
4. C. Raffel, N. Shazeer et al., “Exploring the limits of transfer learning with a unified text-to-text transformer,” Journal of Machine Learning Research, 2020.
5. N. Houlsby, A. Giurgiu et al., “Parameter-efficient transfer learning for nlp,” in International Conference on Machine Learning (ICML), 2019.
6. B. Lester, R. Al-Rfou, and N. Constant, “The power of scale for parameter-efficient prompt tuning,” in Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021.
7. E. J. Hu, Y. Shen et al., “Lora: Low-rank adaptation of large language models,” in International Conference on Learning Representations (ICLR), 2022.
8. N. Ding et al., “Parameter-efficient fine-tuning of large language models: A survey,” arXiv preprint arXiv:2303.15647, 2023.
9. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal,
10. Ku¨ttler, M. Lewis, W.-t. Yih, T. Rockta¨schel, S. Riedel, and D. Kiela, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” in Advances in Neural Information Processing Systems (NeurIPS), 2020.
11. K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, “Realm: Retrieval-augmented language model pre-training,” in Proceedings of the 37th International Conference on Machine Learning (ICML), 2020.

12. V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," in Proceedings of EMNLP, 2020.
13. G. Izacard, E. Grave et al., "Atlas: Few-shot learning with retrieval augmented language models," in International Conference on Machine Learning (ICML), 2022.
14. S. Borgeaud, A. Mensch et al., "Improving language models by retriev- ing from trillions of tokens," Proceedings of ICML, 2022.
15. W. Shi et al., "Replug: Retrieval-augmented black-box language mod- els," arXiv preprint arXiv:2301.12652, 2023.
16. Y. Gao et al., "Retrieval-augmented generation for large language models: A survey," arXiv preprint arXiv:2312.10997, 2024.
17. T. Schick et al., "Toolformer: Language models can teach themselves to use tools," Advances in Neural Information Processing Systems (NeurIPS), 2023.
18. R. Nakano et al., "Webgpt: Browser-assisted question-answering with human feedback," in arXiv preprint arXiv:2112.09332, 2022.
19. Y. Zhang et al., "A survey on hallucination in large language models,"
20. ACM Computing Surveys, 2023.
21. P. Manakul et al., "Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models," in EMNLP, 2023.
22. A. Q. Jiang, A. Sablayrolles, A. Mensch et al., "Mistral 7b," arXiv preprint arXiv:2310.06825, 2023.
23. N. Thakur et al., "Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models," Proceedings of NeurIPS, 2023.
24. Q. Jin, B. Dhingra, Z. Liu, W. Cohen, and X. Lu, "Pubmedqa: A dataset for biomedical research question answering," in Proceedings of EMNLP- IJCNLP, 2019.
25. T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh,
26. C. Alberti, D. Epstein, I. Polosukhin, M. Kelcey, J. Devlin, K. Lee et al., "Natural questions: A benchmark for question answering research," Transactions of the Association for Computational Linguistics, vol. 7,pp. 453–466, 2019.