RAG "Hype" vs. Reality Finetuning as the Cornerstone for Deep Knowledge and Long Context in LLMs*

Ashioya Jotham Victor† ASHIOYA@KABARAK.AC.KE *Kenya*

Abstract

Large Language Models (LLMs) require mechanisms to integrate external, specific, and up-to-date knowledge beyond their static pre-training data. Retrieval-Augmented Generation (RAG) and finetuning represent two dominant paradigms to address this, but their fundamental capabilities and long-term viability warrant critical evaluation. This position paper argues that RAG, while offering practical utility for accessing dynamic information and mitigating hallucination, constitutes a potentially overhyped approach with significant inherent limitations fundamentally tied to its reliance on discrete retrieval steps. We contend that RAG's effectiveness is bottlenecked by retrieval quality, often leads to superficial knowledge integration, struggles with complex reasoning requiring synthesis across information pieces, and faces challenges in robustly leveraging long context windows. Furthermore, the focus on auxiliary technologies like vector databases within the RAG ecosystem can distract from core model capabilities. Conversely, we argue that finetuning, by directly modifying the model's parameters, enables deeper, more nuanced assimilation of domain knowledge and task-specific skills. This parametric adaptation provides a more robust foundation for complex reasoning and is crucial for unlocking true long-context understanding and utilization within the model itself. While acknowledging finetuning's computational and data requirements, we conclude that it offers a more powerful and durable pathway towards developing truly specialized, knowledgeable, and context-aware LLMs, positioning it as the cornerstone for advancing LLM capabilities beyond the architectural constraints of current RAG systems.

Keywords: LLMs, RAG, Vector Databases, Finetuning

1. Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities across a spectrum of natural language tasks, driven by advancements in scale, architecture (primarily the Transformer), and pre-training on vast corpora (Zhao et al., 2023). Influential models like GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2022), and LLaMA (Touvron et al., 2023) exemplify this progress. Their ability to generate coherent text, translate languages, and answer questions has positioned them as transformative technologies. However, a fundamental limitation persists: the knowledge encoded within their parameters is inherently static, reflecting the data cutoff of their pre-training phase (Cheng et al., 2024). Furthermore, LLMs can struggle with domain-specific knowledge and reasoning over complex information not explicitly seen during training and can be prone to generating plausible but factually incorrect statements, commonly termed 'hallucinations' (Zhao et al., 2023; Gao

* This paper was by inspired by this tweet: https://x.com/jobergum/status/1872923872007217309

† Kabarak University

et al., 2024; Lialin et al., 2023). To overcome these limitations and enhance LLM utility in real-world applications requiring current, specialized, or proprietary information, two primary strategies have emerged: Retrieval-Augmented Generation (RAG) (Guu et al., 2020; Lewis et al., 2020) and finetuning (Parthasarathy et al., 2024). RAG systems dynamically supplement the model's internal knowledge by retrieving relevant text snippets from an external corpus (often indexed in vector databases) and providing these snippets as context during the generation process (Gao et al., 2024). Finetuning, in contrast, involves further training a pre-trained model on a smaller, task-specific or domain-specific dataset, directly adapting the model's internal parameters to imbibe new knowledge or skills (Zhang et al., 2023). This includes both full parameter updates and more recent parameter-efficient finetuning (PEFT) techniques like LoRA (Hu et al., 2021; Lialin et al., 2023; Ding et al., 2023). While both approaches aim to improve LLM knowledge and performance, they operate via fundamentally different mechanisms, leading to distinct trade-offs and, we argue, significantly different long-term potential. Current discourse often presents RAG as a flexible and efficient solution, sometimes accompanied by considerable hype, particularly concerning the associated infrastructure like vector databases (Wiggers, 2024). This paper challenges that perspective. We posit that RAG, despite its utility in specific scenarios, represents a comparatively shallow form of knowledge integration, inherently constrained by the efficacy and limitations of its retrieval component. Its perceived advantages often mask fundamental weaknesses in handling complex reasoning and achieving true long-context understanding.

Conversely, we argue that finetuning, by modifying the parametric knowledge base of the LLM itself, offers a more robust and powerful pathway towards deep knowledge assimilation, nuanced domain adaptation, and the effective utilization of long-context capabilities. While finetuning presents its own challenges, particularly regarding computational resources and data curation, we contend that it provides the necessary foundation for building models capable of genuine expertise and sophisticated reasoning within specific domains. Therefore, this paper advocates for the primacy of finetuning as the cornerstone strategy for developing advanced, knowledgeable LLMs, particularly for complex tasks demanding deep understanding and robust long-context processing, viewing RAG more as a supplementary tool or potentially a temporary solution whose necessity may diminish as core model capabilities improve.

This position paper is structured as follows: Section 2 delves into the mechanisms of RAG, critically examining its limitations, including retrieval dependency, superficial knowledge integration, and challenges with long context, alongside a critique of the vector database hype. Section 3 explores finetuning techniques, arguing for their superiority in achieving deep knowledge integration and enabling effective long-context reasoning. Section 4 provides a comparative analysis, discussing trade-offs, hybrid approaches, and reinforcing the core thesis. Finally, Section 5 concludes by summarizing the arguments and reiterating the case for finetuning as the more promising direction for future LLM development in complex application domains.

2. Retrieval-Augmented Generation: Mechanisms and Critiques

Retrieval-Augmented Generation (RAG) has gained significant traction as a method to ground LLM outputs in external knowledge, thereby potentially reducing hallucinations

and incorporating up-to-date information (Lewis et al., 2021; Chen et al., 2023a). However, a closer examination reveals fundamental limitations that challenge its perception as a robust, long-term solution for deep knowledge integration and complex reasoning.

2.1. The RAG Mechanism

At its core, the standard RAG pipeline involves intercepting a user query, using it to search an external knowledge corpus, retrieving relevant information chunks, and then feeding these chunks alongside the original query as augmented context to an LLM for generation (Gao et al., 2024). RAG is like giving a librarian a question to find relevant books, then having a writer use those books to answer. It pulls information from external sources, such as company documents, to make LLM responses more accurate and up-to-date, especially for things like recent news or specific business data (Schreiner, 2023). This process typically relies on several key components:

(i) an external knowledge base (e.g., documents, databases);

(ii) an indexing mechanism, frequently employing dense vector embeddings stored in specialized vector databases, to facilitate semantic search (Gao et al., 2024)

(iii) a retriever module that searches the index based on the query and selects relevant context chunks; and

(iv) a generator LLM that synthesizes the final response based on the query and the retrieved context.

2.2. The Retrieval Bottleneck: RAG's Fundamental Constraint

The most significant vulnerability of the RAG architecture lies in its absolute dependence on the retrieval step (Ji et al., 2023). The quality of the final generation is fundamentally capped by the quality and relevance of the retrieved information. Failure modes in retrieval are numerous and critical:

(i) Missing Information: The correct information might not exist in the external knowledge base (Barnett et al., 2024)

(ii) Retrieval Failure (Recall): Relevant documents exist but are not retrieved by the system.

(iii) Irrelevant Retrieval (Precision): Documents are retrieved but do not actually contain the necessary information or are misaligned with the query's true intent.

(iv) Information Distillation Failure: The relevant information is present within the retrieved chunks, but the LLM fails to identify, extract, or synthesize it correctly, potentially distracted by irrelevant surrounding text.

(v) Noise and Contradiction: Retrieved chunks may contain conflicting, inaccurate, or noisy information, which can mislead the generator LLM.

These potential failures make RAG systems inherently fragile; performance is inextricably tied to the comprehensiveness and cleanliness of the knowledge source, the effectiveness of the chunking strategy, the quality of the embeddings, and the capability of the retriever model.

2.3. RAG and the Long Context Illusion

While LLMs with longer context windows are emerging, using RAG to simply "fill" this window with more retrieved chunks is not a panacea for long-context reasoning (Liu et al., 2023; Ovadia et al., 2024). The Databricks blog "Long Context RAG Performance of LLMs" shows that performance decreases after certain context sizes (e.g., 32k tokens for Llama-3.1-405b, 64k for GPT-4-0125-preview), with only recent models like GPT-4o maintaining consistency (Chase and Adebayo, 2023). The "lost in the middle" effect, where LLMs prioritize information at the beginning or end of long contexts, further complicates RAG's utility (Petrosyan, 2024). Additionally, retrieving more documents increases noise, as shown by recall saturation at 96k tokens for some datasets, suggesting that simply adding more chunks does not guarantee improved performance. Several issues arise:

(i) Attention Limitations: Many LLMs exhibit difficulty in effectively utilizing information spread across very long contexts, often suffering from a "lost in the middle" effect where information at the beginning or end is weighted more heavily than information in the middle. RAG does not intrinsically solve this; it merely provides the long context that the base LLM might struggle to process effectively.

(ii) Retrieval Quality Degradation: As the target context grows (either longer documents or more documents), identifying the truly salient chunks becomes exponentially harder, increasing the likelihood of retrieving noise or missing critical pieces.

(iii) Performance Saturation/Degradation: Studies suggest that simply increasing the number of retrieved chunks does not always improve, and can sometimes even degrade, RAG performance, potentially due to increased noise or distraction for the generator. Recent advancements in native long-context models are beginning to challenge RAG's dominance, even in retrieval-centric tasks.

2.4. Deconstructing the Vector Database Hype

Vector databases have become closely associated with RAG, often positioned as a core enabling technology. The DEV Community article "How About Ditching the Hype: Do We Really Need a Specialized Vector Database?" questions the necessity of specialized vector databases, noting that traditional databases are incorporating vector search capabilities, potentially reducing the need for separate systems (Singh, 2024). The Substack post "Vector Database is not a separate database category" predicts that specialized vector databases will lose momentum as integrated solutions gain traction (Morgan, 2023). While useful for efficient similarity search on high-dimensional embeddings, their role and importance can be overstated(Bergum, 2024):

(i) Component, Not Solution: They are merely one part of the retrieval pipeline. Their effectiveness is entirely dependent on the quality of the upstream embedding model

and the appropriateness of the chosen similarity metric (e.g., cosine similarity) for capturing true semantic relevance, which is not always guaranteed (AlShikh, 2023).

(ii) Garbage In, Garbage Out: A vector database cannot compensate for poor data quality, inaccurate information, or biases present in the source documents.

(iii) Distraction from Fundamentals: The intense focus on vector database technology can distract from more fundamental challenges in RAG, such as optimal chunking strategies, retriever training, and handling complex query logic. Simpler retrieval methods (e.g., keyword search, hybrid approaches) might be sufficient or even superior in some contexts. The perceived necessity of specialized vector databases is also being questioned as traditional databases increasingly incorporate vector capabilities.

2.5. RAG as a Tactical Solution

In summary, while RAG provides a valuable mechanism for accessing dynamic external facts and attributing sources (Lewis et al., 2020; Balaguer et al., 2024), its architectural limitations – the retrieval bottleneck, superficial knowledge integration, challenges in scaling effectively with context length, added latency, and operational overhead – suggest it is not the foundational solution for building deeply knowledgeable and capable LLMs. Several researchers and practitioners view RAG as a clever workaround for the static knowledge and finite context limitations of earlier models (Gao et al., 2024; Xu et al., 2024),a tactical tool rather than a strategic long-term direction for achieving genuine understanding and complex reasoning abilities. Its strengths lie in scenarios demanding access to rapidly changing data or verifiable sourcing, but its limitations hinder its applicability for tasks requiring deep, integrated expertise.

3. Finetuning for Deep Integration and Long Context

While RAG offers a way to augment LLMs with external information at inference time, finetuning provides a mechanism to fundamentally alter the model's internal knowledge and capabilities through continued training on specialized data. This parametric adaptation, we argue, leads to deeper knowledge integration, more nuanced behavioural changes, and is ultimately essential for effectively leveraging long context windows for complex tasks.

3.1. Mechanisms of Finetuning

Finetuning involves updating the weights of a pre-trained LLM using a smaller, curated dataset relevant to a specific domain or task (Parthasarathy et al., 2024). Two main approaches exist:

(i) Full Finetuning: All (or a significant portion) of the model's parameters are updated during the training process. This allows for substantial adaptation but incurs significant computational cost and requires careful management to avoid "catastrophic forgetting" of the model's general capabilities (Brown et al., 2020; McCloskey and Cohen, 1989; Luo et al., 2025; Kirkpatrick et al., 2017).

(ii) Parameter-Efficient Fine-Tuning (PEFT): These techniques aim to reduce the computational burden and memory requirements of finetuning by updating only a small subset of the model's parameters or by introducing a small number of new, trainable parameters(Ding et al., 2023; Lialin et al., 2023). Methods like Low-Rank Adaptation (LoRA)(Hu et al., 2021), adapters, and prompt tuning fall under this category. PEFT makes finetuning more accessible and efficient, often achieving performance comparable to full finetuning on specific tasks.

3.2. Achieving Deep Knowledge and Skill Integration

Unlike RAG, which treats the LLM largely as a fixed processor acting on retrieved context, finetuning directly modifies the model's internal representation of knowledge (Balaguer et al., 2024). This process allows for:

(i) Assimilation of Domain-Specific Knowledge: Finetuning on domain texts (e.g., medical research, legal documents, specific coding practices) embeds relevant terminology, concepts, and relationships into the model's parameters. The model doesn't just see the terms; it learns their usage patterns and connections.

(ii) Adaptation of Style and Behavior: Finetuning can reliably shape the model's output style, tone, persona, and adherence to specific formats or reasoning patterns. This is crucial for applications requiring consistent branding, specific interaction protocols, or complex instruction following (Zhang et al., 2023).

(iii) Improved Reasoning within a Domain: By learning from examples within a specific domain, finetuned models can potentially improve their ability to perform multi-step reasoning, inference, and synthesis using the learned domain knowledge, rather than solely relying on retrieving explicit facts (Brokman and Kavuluru, 2024).

This deep integration contrasts sharply with RAG's reliance on potentially noisy or incomplete retrieved snippets, enabling finetuned models to exhibit more consistent, nuanced, and reliable behavior on their specialized tasks (Anisuzzaman et al., 2025).

3.3. Addressing the Challenges of Finetuning

It is important to acknowledge the challenges associated with finetuning. Full finetuning demands substantial computational resources (GPUs, time), significant amounts of high-quality training data (which may require expensive human labeling), and expertise in training methodologies to prevent issues like catastrophic forgetting (Zhang et al., 2023; Luo et al., 2025; Gutta, 2023). However, several factors mitigate these concerns:

(i) PEFT Efficiency: Techniques like LoRA drastically reduce the computational and memory footprint, making finetuning feasible even with moderate resources(Hu et al., 2021; Liu et al., 2024; Dettmers et al., 2023).

(ii) Data Curation Efforts: While data quality is paramount, ongoing research into data selection, synthetic data generation, and unsupervised/semi-supervised finetuning techniques aims to reduce the reliance on massive labeled datasets

(iii) Engineering vs. Architecture: These challenges, while significant, are largely engineering and resource problems. They concern the process of adapting the model. In contrast, RAG's limitations, particularly the retrieval bottleneck, are arguably more fundamental architectural constraints (Ovadia et al., 2024).

Therefore, while the costs and complexities of finetuning are real, they represent investments in building a more deeply capable and adapted model, rather than relying on an external retrieval mechanism with inherent fragility. Essentially, finetuning offers a powerful mechanism for embedding deep domain knowledge, shaping model behavior, and crucially, enabling effective utilization of long context windows. By directly modifying the model's parameters, it achieves a level of integration and adaptation that RAG, by its nature, cannot match, making it the cornerstone for developing truly specialized and contextually adept LLMs.

4. Discussion

While both aim to bridge the gap left by static pre-training, their underlying mechanisms lead to fundamental differences in performance, robustness, and suitability for various tasks.

4.1. Valid Niches for RAG

Despite the critiques presented, RAG remains a valuable tool in specific contexts. Its strength lies in scenarios where:

(i Information is Highly Dynamic: For applications requiring access to constantly changing information (e.g., news, stock prices, real-time inventory), RAG's ability to query updated knowledge bases without retraining is crucial (Cliver, 2024).

(ii Verifiability is Paramount: When users need to trace answers back to specific source documents (e.g., legal research, customer support referencing manuals), RAG's retrieval mechanism provides inherent citability (Choudhary, 2024).

(iii Knowledge Base is Vast and Explicit: When dealing with enormous, structured or semi-structured knowledge bases where explicit fact retrieval is the primary goal, RAG can be effective (Logic20/20).

However, acknowledging these niches does not negate the argument that RAG is fundamentally limited for tasks requiring deep understanding, complex reasoning, or nuanced behavioural adaptation. It serves best as an information provision mechanism, not a knowledge integration one.

4.2. Finetuning: The Key to Mastering Long Context

The advent of LLMs with increasingly long context windows (e.g., hundreds of thousands or even millions of tokens) presents new opportunities and challenges. While RAG can provide information to fill these windows, finetuning is arguably essential to enable the model to effectively process and reason over such extended contexts (Beltagy et al., 2020; Pan, 2023; Petrosyan, 2024; Chen et al., 2023b; Gao et al., 2025).

(i) Adapting Attention Mechanisms: Effective utilization of long context likely requires models to adapt their internal mechanisms, such as attention patterns, to better track dependencies and relationships over vast spans of text. Finetuning on tasks requiring long-range understanding can potentially induce these adaptations.

(ii) Learning Long-Range Dependencies: Specific finetuning strategies and datasets can be designed to explicitly teach the model to identify and utilize information spread across long documents or dialogue histories. This goes beyond the capability of RAG, which merely presents the information without guaranteeing the model can effectively connect distant pieces.

(iii) Task-Specific Context Utilization: Even with large base model context windows, finetuning may be required to teach the model how to best use that context for a specific downstream task (e.g., summarizing a long legal document vs. answering specific questions about it). Studies are showing that finetuning can significantly boost performance on long-context benchmarks compared to relying solely on the base model's capabilities.

Thus, while RAG might seem like a shortcut to leveraging external data in long contexts, finetuning offers a more fundamental path to improving the model's intrinsic ability to handle and reason over long sequences effectively.

4.3. The Rise of Hybrid Approaches

The emergence of hybrid models that combine RAG and finetuning further underscores the limitations of RAG operating in isolation. Techniques like Retrieval-Augmented Fine-Tuning (RAFT) (Tian et al., 2024), or approaches that finetune the retriever and/or the generator specifically for the RAG task (Prajna AI Wisdom, 2024; WolframRavenwolf, 2024; Choudhury, 2024), demonstrate a recognition that optimal performance often requires both parametric adaptation and dynamic retrieval. While these hybrids can be powerful, their existence implicitly supports our thesis: RAG alone is often insufficient for complex tasks, requiring the deeper integration provided by finetuning to reach peak performance. The complexity of designing and tuning these hybrid systems also moves away from the initial appeal of RAG as a simpler plug-and-play solution.

5. Conclusion

Equipping LLMs with specific, up-to-date, and deeply integrated knowledge is crucial. While RAG offers utility for dynamic information access and attribution, its perception as a foundational solution is inflated, masking architectural limitations like retrieval dependency and superficial knowledge integration.

Conversely, finetuning offers a robust path to genuine domain expertise and effective long-context reasoning by parametrically adapting models. While RAG has its niches and hybrid systems show promise, finetuning remains the cornerstone for building truly specialized, knowledgeable, and contextually adept LLMs. Future advancements in complex, knowledge-intensive LLM applications will likely rely on continued progress in finetuning methodologies and core model capabilities.

References

Waseem AlShikh. RAG vs. vector database: What's the difference? Writer Engineering Blog, October 2023. URL https://writer.com/engineering/rag-vector-database/. Accessed: 2025-04-14.

D.M. Anisuzzaman, Jeffrey G. Malins, Paul A. Friedman, and Zachi I. Attia. Fine-tuning large language models for specialized use cases. *Mayo Clinic Proceedings: Digital Health*, 3(1):100184, 2025. ISSN 2949-7612. doi: https://doi.org/10.1016/j.mcpdig.2024.11.005. URL https://www.sciencedirect.com/science/article/pii/S2949761224001147.

Angels Balaguer, Vinamra Benara, Renato Luiz de Freitas Cunha, Roberto de M. Estevão Filho, Todd Hendry, Daniel Holstein, Jennifer Marsman, Nick Mecklenburg, Sara Malvar, Leonardo O. Nunes, Rafael Padilha, Morris Sharp, Bruno Silva, Swati Sharma, Vijay Aski, and Ranveer Chandra. Rag vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture, 2024. URL https://arxiv.org/abs/2401.08406.

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. Seven failure points when engineering a retrieval augmented generation system, 2024. URL https://arxiv.org/abs/2401.05856.

Iz Beltagy, Kyle Lo, and Arman Cohan. Longformer: The long-document transformer, 2020. URL https://arxiv.org/abs/2004.05150.

Jo Bergum. Post discussing RAG vs Finetuning concepts. Post on X.com (formerly Twitter), December 2024. URL https://x.com/jobergum/status/1872923872007217309. Accessed: 2025-04-14. Handle: @jobergum.

Aviv Brokman and Ramakanth Kavuluru. How important is domain specificity in language models and instruction finetuning for biomedical relation extraction?, 2024. URL https://arxiv.org/abs/2402.13470.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Greg Chase and Samuel Adebayo. Long-context LLMs struggle with RAG: Introducing Focused Transformation and Summarization to improve performance. Databricks Blog, November 2023. URL https://www.databricks.com/blog/long-context-rag-performance-llms. Accessed: 2025-04-14.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation, 2023a. URL https://arxiv.org/abs/2309.01431.

Shouyuan Chen, Sherman Wong, Liang Chen, and Yuandong Tian. Extending context window of large language models via position interpolation, 2023b. URL https://arxiv.org/abs/2306.15595.

Jeffrey Cheng, Marc Marone, Orion Weller, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. Dated data: Tracing knowledge cutoffs in large language models, 2024. URL https://arxiv.org/abs/2403.12958.

Gautam Choudhary. Understanding source attribution in RAG, January 2024. URL https://gautam75.medium.com/understanding-source-attribution-in-rag-6c8d64cfaed8. Accessed: 2025-04-14.

Bedabrata Choudhury. Adapting RAG & finetuning through intelligent document processing. LinkedIn Pulse Article, March 2024. URL https://www.linkedin.com/pulse/adapting-rag-finetuning-through-intelligent-bedabrata-choudhury-pl58c. Accessed: 2025-04-14.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Green, Jason Zhou, Stephen Gaffney, Andres Garcia, Noah Fiedel, Mostafa Dehghani, Azalia Mirhoseini, Barret Zoph, Jan A Botha, Brian Jennison, Monsi Casado, Jonathan FitzGerald, Yongwoo Seol, Edgar Vassilvitckii, Corey Folly, Pierre Crepy, Charline Nogueira, Shuo-Yiin Wu, Thomas Kipf, Alex Karp, Mark Van Der Werf, Machel Diaz, Nate Kushman, Slav Petrov, William Fedus, Jiaming Liu, Zirui Wang, Yinfei Hsieh, John Hawkins, Mohsin Mohiuddin, Mi Li, Ilya Sutskever, Samy Bengio, Jeff Dean, and Noah Fiedel. PaLM: Scaling language modeling with pathways, 2022. URL https://arxiv.org/abs/2204.02311.

Fredric Cliver. Dynamic Retrieval-Augmented Generation: Real-Time information update for enhanced conversational experiences. Medium, March 2024. URL https://fredriccliver.medium.com/dynamic-retrieval-augmented-generation-real-time-information-update-for-enhanced-conversational-1bd83bcebe32. Accessed: 2025-04-14.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023. URL https://arxiv.org/abs/2305.14314.

Ning Ding, Yujia Chen, Bowen Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Lin, and Zhiyuan Liu. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023. doi: 10.1038/s42256-023-00627-4.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. How to train long-context language models (effectively), 2025. URL https://arxiv.org/abs/2410.02660.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Sun. Retrieval-augmented generation for large language models: A survey, 2024. URL https://arxiv.org/abs/2312.10997.

Sreedhar Gutta. Fine-tuning large language models: A comprehensive guide. Analytics Vidhya Blog, August 2023. URL https://www.analyticsvidhya.com/blog/2023/08/fine-tuning-large-language-models/. Accessed: 2025-04-14.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training, 2020. URL https://arxiv.org/abs/2002.08909.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large-scale language models, 2021. URL https://arxiv.org/abs/2106.09685.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys (CSUR)*, 55(12):1–38, 2023. doi: 10.1145/3571730.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, March 2017. ISSN 1091-6490. doi: 10.1073/pnas.1611835114. URL http://dx.doi.org/10.1073/pnas.1611835114.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2020. URL https://arxiv.org/abs/2005.11401.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL https://arxiv.org/abs/2005.11401.

Vladislav Lialin, Vijeta Deshpande, and Anna Rumshisky. Scaling down to scale up: A guide to parameter-efficient fine-tuning, 2023. URL https://arxiv.org/abs/2303.15647.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts, 2023. URL https://arxiv.org/abs/2307.03172.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation, 2024. URL https://arxiv.org/abs/2402.09353.

Logic20/20. Enhancing knowledge base interactions with RAG architecture. Logic20/20 Insight. URL https://logic2020.com/insight/enhancing-knowledge-base-interactions-with-rag-architecture/. Accessed: 2025-04-14.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning, 2025. URL https://arxiv.org/abs/2308.08747.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press, 1989. doi: https://doi.org/10.1016/S0079-7421(08)60536-8. URL https://www.sciencedirect.com/science/article/pii/S0079742108605368.

Timothy Prickett Morgan. Vector database is not a separate market. NextWord Substack Newsletter, December 2023. URL https://nextword.substack.com/p/vector-database-is-not-a-separate. Accessed: 2025-04-14.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. Fine-tuning or retrieval? comparing knowledge injection in llms, 2024. URL https://arxiv.org/abs/2312.05934.

Wenbo Pan. Long context fine-tuning: a how-to guide using LoRA and Axolotl. Hugging Face Blog, October 2023. URL https://huggingface.co/blog/wenbopan/long-context-fine-tuning. Accessed: 2025-04-14.

Venkatesh Balavadhani Parthasarathy, Ahtsham Zafar, Aafaq Khan, and Arsalan Shahid. The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities, 2024. URL https://arxiv.org/abs/2408.13296.

Vahan Petrosyan. RAG vs. fine-tuning vs. long-context LLMs. SuperAnnotate Blog, January 2024. URL https://www.superannotate.com/blog/rag-vs-long-context-llms. Accessed: 2025-04-11.

Prajna AI Wisdom. Hybrid approaches: Combining RAG and finetuning for optimal LLM performance. Medium, March 2024. URL https://prajnaaiwisdom.medium.com/hybrid-approaches-combining-rag-and-finetuning-for-optimal-llm-performance-35d2bf3582a9. Accessed: 2025-04-14.

Erik Schreiner. What is retrieval-augmented generation? NVIDIA Blog, August 2023. URL https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/. Accessed: 2025-04-14.

Gaurav Singh. How about ditching the hype: Do we really need a specialized vector database? Dev.to, February 2024. URL https://dev.to/gaurav274/how-about-ditching-the-hype-do-we-really-need-a-specialized-vector-database-3241. Accessed: 2025-04-14.

Kang Tian, Hong Pu, Rui Zhang, Weihua Lin, Yuanbin Guo, Ya-Fei Zhou, Jian-Guang Liu, and Hui Zhang. Raft: Retrieval augmented fine-tuning for accuracy and plug-and-play enhancement, 2024. URL https://arxiv.org/abs/2403.10131.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Gan, Vedanuj Gandra, Lakshya Goyal, Elena Hartshorn, Sefa Hosseini, Safa Hosseini, Gautier Izacard, Zian Jiang, Wenliang Ju, Bill James, Marc-Alexandre Lachaux, Thibaut Lacroix, Thomas Lavril, Jenya Lee, Guillaume Lample, Xavier Lefrancois, Victor Logatskiy, Tamar A Mandelbaum, Josephine Masson, Clement Jean Mathis, John Miller, Ivan Molybog, Ylan Monchiero, Vedanuj Murthy, Anthony Nguyen, Naman Nikolai, Sergey Panteleev, Antonio Penedo, Andrew Poulton, Jeremy Pytlarz, Thomas Rizzett, Andrew Roberts, Richard Rogel, Baptiste Roziere, Julian Michael Rusch, Julia Seznec, Teven Le Scao, Justine Schulte, Anush Selvan, Adi Sharma, Egor Shleifer, Rui Silva, Serge Miguel Sonnenburg, Peter Stock, Hongyu Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narayanan, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. LLaMA: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Kyle Wiggers. Why vector databases are having a moment as the AI hype cycle peaks. TechCrunch, April 2024. URL https://techcrunch.com/2024/04/20/why-vector-databases-are-having-a-moment-as-the-ai-hype-cycle-peaks/. Accessed: 2025-04-14.

WolframRavenwolf. Hybrid RAG and finetuning systems. Discussion thread on Reddit r/LocalLLaMA, January 2024. URL https://www.reddit.com/r/LocalLLaMA/comments/19ctl5b/hybrid_rag_and_finetuning_systems/. Accessed: 2025-04-14.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. Retrieval meets long context large language models, 2024. URL https://arxiv.org/abs/2310.03025.

Zheng Zhang, Aston Zhang, Mu Li, Hai Zhao Lin, Sheng Yang, Tianlong Wang, Xiang Chen, Qipeng Gao, Shoumik Shah, Yan Lv, Kuang-Hua Chen, Kevin He, Zhiheng Shou, Xuanjing Ma, and Kang Chen. Instruction tuning for large language models: A survey, 2023. URL https://arxiv.org/abs/2308.10792.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junkun Zhang, Zican Dong, Yifan Du, Chengrui Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2023. URL https://arxiv.org/abs/2303.18223.