# Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely

*S. Zhao, Y. Yang, Z. Wang, Z. He, L. K. Qiu, and L. Qiu, “Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey  on How to Make your LLMs use External Data More Wisely,” arXiv (Cornell University), Sep. 2024, doi: 10.48550/arxiv.2409.14924.*

This paper discusses the integration of external data into large language models to boost their performance in real-world tasks, focusing on techniques like Retrieval-Augmented Generation (RAG) and fine-tuning. Although these approaches display great promise, deploying data augmented LLMs in specialized fields face significant challenges such as retrieving relevant data, interpreting user intent, and leveraging reasoning capabilities for complex tasks.The authors argue that there is no universal solution since under performance often stems from misidentifying the core focus of the task or the need to address multiple intertwined capabilities.

The authors implement a task categorization framework for RAG, classifying user queries into four levels based on the type of external data used and task focus: explicit fact queries, implicit fact queries, interpretable rationale queries, and hidden rationale queries. They define, provide datasets and information on challenges and good practices for each query type, as well as explore three methods of integrating external data into LLMs—context-based integration, small model augmentation, and fine-tuning—pointing out their strengths, weaknesses, and best uses. It is a guide on how to understand the needs of data, identify bottlenecks, and systematically develop robust LLM applications.

**Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks**

*P. S. H. Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP tasks,” Neural Information Processing Systems, vol. 33, pp. 9459–9474, May 2020, [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf*

This paper develops Retrieval-Augmented Generation (RAG) models that are improvements over pre-trained models. They combine parametric memory (in the form of stored within the model) with non-parametric memory (external sources that can be used, e.g., Wikipedia). There are key limitations of such pre-trained models, primarily including difficulties in accessing and manipulating and updating knowledge but also challenges in providing a provenance for their decisions.  
The authors introduce a general-purpose fine-tuning approach for RAG models and evaluate them on knowledge-intensive NLP tasks. In their framework, the parametric memory is a pre-trained seq2seq model, while the non-parametric memory is a dense vector index of Wikipedia accessed via a pre-trained neural retriever. Two RAG formulations are compared:

1. Sequence-level retrieval, where the same retrieved passages are used across the entire generated sequence.

2. Token-level retrieval, where the retrieval for each token is potentially a different passage.

Main takeaway points are:

-State-of-the-art performance for RAG models on three open-domain question-answering tasks over parametric seq2seq models as well as task-specific retrieve-and-extract architectures.

-For language generation, the RAG model generated more specific, diverse and factual language than a state-of-the-art parametric seq2seq baseline.

The potential of RAG models in overcoming the limits of knowledge representation and retrieval is well demonstrated by combining pre-trained parametric models with non-parametric memory.

**LoRA Land: 310 Fine-tuned LLMs that Rival GPT-4, A Technical Report**

*J. Zhao et al., “LoRA Land: 310 Fine-tuned LLMs that Rival GPT-4, A Technical Report,” arXiv (Cornell University), Apr. 2024, doi: 10.48550/arxiv.2405.00732.*

This paper explores how Low Rank Adaptation (LoRA) can be used for fine-tuning large language models (LLMs) without using too many parameters. LoRA cuts down on the number of trainable parameters and memory needed, but still keeps up with the performance you'd get from full fine-tuning. The paper checks out how LoRA can be used in real life, looks at how good the models are, how well they fine-tune, and how efficient they are when deployed.

The authors put 310 LoRA fine-tuned models to the test over 10 base models and 31 tasks. On average, these models improved by 34 points over the base models and by 10 points over GPT-4 when using 4-bit LoRA adapters. So, even if there is a shortage of resources, LoRA can still prove capable. The paper explores which base models are best for fine-tuning and discusses some tricks for guessing how well fine-tuning will go based on how tough the task is.

LoRAX: Efficient Inference for LoRA Models:

LoRAX, an open-source multi-LoRA inference server that lets you run multiple fine-tuned models on a single GPU. It makes smart use of shared base model weights and dynamic adapter loading to get higher efficiency at a lower cost.

LoRA Land:  
This is a web application powered by LoRAX that hosts 25 LoRA fine-tuned Mistral-7B models on just one NVIDIA A100 GPU with 80GB memory. It shows off how you can save money and get value by deploying several specialized models instead of relying on one all-purpose LLM.

The results show that LoRA is extremely capable of fine-tuning LLMs in real-world situations, finding a balance between performance and efficiency. It's especially useful when you need to roll out multiple specialized models without going overboard on computing power.

# Dial-insight: Fine-tuning Large Language Models with High-Quality Domain-Specific Data Preventing Capability Collapse

*J. Sun et al., “Dial-insight: Fine-tuning Large Language Models with High-Quality  Domain-Specific Data Preventing Capability Collapse,” arXiv (Cornell University), Mar. 2024, doi: 10.48550/arxiv.2403.09167.*

This paper investigates a two-stage procedure for fine-tuning huge models to target domain applications, where there are a possibility of loss of generative capabilities. It introduces an affordable multidimensional quality assessment framework that seeks to validate the integrity of this data and explores methods constructing highly diversified prompts and therefore coverage ranges into tasks and expressions of generation. Using a service provider and customer interaction dataset in the real estate industry, the authors demonstrate an impressive correlation between data quality and performance.

Importantly, the findings show that fine-tuning with domain-specific data, generated through this approach, improves domain-specific proficiency without compromising generalization abilities of the model, even if only using domain-specific data. It stresses data diversity and quality as optimally pertinent to the performance optimization on the specialized domains.

# Gemma: Open Models Based on Gemini Research and Technology

*G. Team et al., “GemMa: Open models based on Gemini research and technology,” arXiv (Cornell University), Mar. 2024, doi: 10.48550/arxiv.2403.08295.*

This paper introduces Gemma, a family of light, state-of-the-art open models based on the research and technology behind the Gemini models. Gemma models lead in language understanding, reasoning, and safety by demonstrating outstanding performance across academic benchmarks. Two model sizes are available today: 2 billion and 7 billion parameters; pretrained and fine-tuned checkpoints are available for both sizes. Gemma outperforms peer open models of similar size on 11 of 18 text-based tasks, with thorough assessments offered on safety, accountability, and model development. The authors conclude that the responsible deployment of such models is critical to advancing safety in advanced models and accelerating innovation in the language modeling space.

# Comparative Analysis of Different Efficient Fine Tuning Methods of Large Language Models (LLMs) in Low-Resource Setting

*K. P. V. Srinivasan, P. Gumpena, M. Yattapu, and V. H. Brahmbhatt, “Comparative analysis of different Efficient Fine Tuning Methods of Large Language Models (LLMS) in Low-Resource Setting,” arXiv (Cornell University), May 2024, doi: 10.48550/arxiv.2405.13181.*

This paper investigates various fine-tuning strategies for large language models (LLMs), building on the findings of Mosbach et al. (2023), who showed that few-shot full-model fine-tuning (Vanilla Fine-Tuning and Pattern-Based Fine-Tuning) and In-Context Learning (ICL) generalize similarly on out-of-domain (OOD) datasets but differ in task adaptation. Given the challenges of memory requirements in these methods, the authors aim to comprehensively compare fine-tuning strategies, including adaptive fine-tuning and LoRA adapters, alongside context distillation as an alternative approach.

The experiments were conducted on two datasets, COLA and MNLI, using state-of-the-art fine-tuning methods. Results indicate that while Vanilla Fine-Tuning and PBFT provide strong OOD generalization, PBFT performs below expectations Vanilla FT on OOD tasks, highlighting the importance of effective prompts. Adaptive fine-tuning and LoRA adapters displays comparable or slightly worse performance than standard fine-tuning, consistent with expectations due to the reduced scope of model adjustments. Notably, context distillation outperformed standard fine-tuning methods, offering a promising alternative.

The findings emphasize that the choice of fine-tuning strategy should depend on the available resources—such as memory, compute, and data—as well as the specific requirements of task adaptability. Context distillation emerges as a particularly effective and resource-efficient method in this study.