SAND2024-14215C

Sandia National Laboratories

CAN SMALLER EXPERT MODULES ENHANCE RAG PERFORMANCE?

Alexander Nemecek, Robert Abbott, Christopher Garasi

Which Complementary Methods Improve LLM Accuracy?

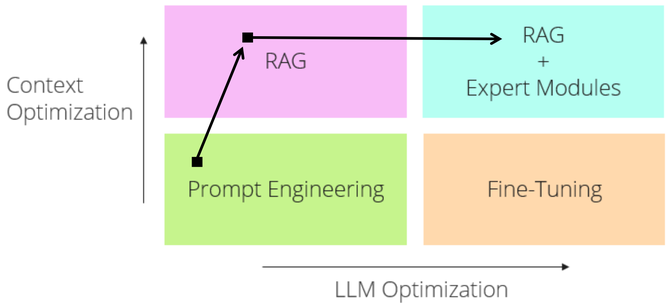


Figure 1: Current optimization timeline for information retrieval from prompt engineering to retrieval augmented generation (RAG) to fine-tuning large language models (LLMs). (Image from https://www.youtube.com/watch?v=ahnGLM-RC1Y)

- Prompt Engineering: Designing language prompts to output the best results from LLMs
- RAG: Information retrieval workflow for external knowledge bases
- Fine-Tuning: Continuing the training process on a smaller, domain-specific dataset to optimize a model for a specific task

How Does a RAG Pipeline Function?

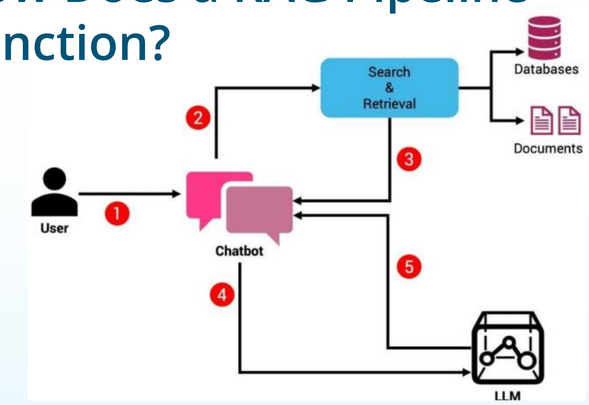
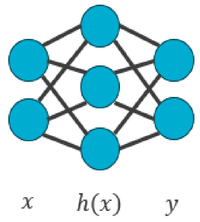


Figure 2: Generalized RAG pipeline for information retrieval. Pipeline follows from the user's input to a search and retrieval framework accessing external information from housed databases and documents, back to the chat interface to supplement the user prompt, to the LLM for generation and then back to the chat interface for a response to the user's initial input. (Image from https://thenewstack.io/freshen-up-llms-with-retrieval-augmented-generation/)

Which Methods are Used to Fine-Tune Expert Modules?

Full Parameter Fine-Tuning

Figure 3: Full parameter fine-tuning, training all model weights from scratch. Given input layer *x*, hidden layer *h(x)*, LLM architecture multiplies *x* by the model weight dimensions, *d* and *k* for *h(x).* Resulting in 1,000,000 trainable parameters.

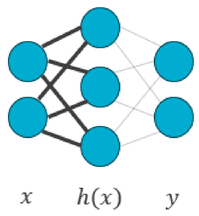


$$h(x) = W_0 x$$

$$d = 1{,}000$$
$$k = 1{,}000$$
$$d * k = 1{,}000 * 1{,}000 = 1{,}000{,}000$$

- Full parameter fine-tuning increases the number of trainable parameters dramatically
- Full parameter fine-tuning requires a significant amount of computational hardware and training time

Parameter Efficient Fine-Tuning (PEFT)



$$h(x) = W_0 x + \Delta W x = W_0 x + BAx$$

$$d = 1{,}000$$
$$k = 1{,}000$$
$$r = 2$$
$$\big((d * r) + (r * k)\big) = \big((1{,}000 * 2) + (2 * 1{,}000)\big) = 4{,}000$$

Figure 4: Parameter efficient fine-tuning via low-rank adaptation (LoRA). Given input layer *x*, hidden layer *h(x)*, LoRA incorporates low-rank matrices given by the weight dimensions, *d* and *k*, and parameter *r* for *h(x).* Resulting in 4, 000 trainable parameters. (Image from https://towardsdatascience.com/fine-tuning-large-language-models-llms-23473d763b91)

- PEFT via LoRA: Low-Rank Adaptation
- LoRA changes the base LLM with a small number of trainable parameters by decomposing the weight matrices
- LoRA fine-tuning occurs with updating ONLY the low-rank matrices

Quantized- LoRA (QLoRA)

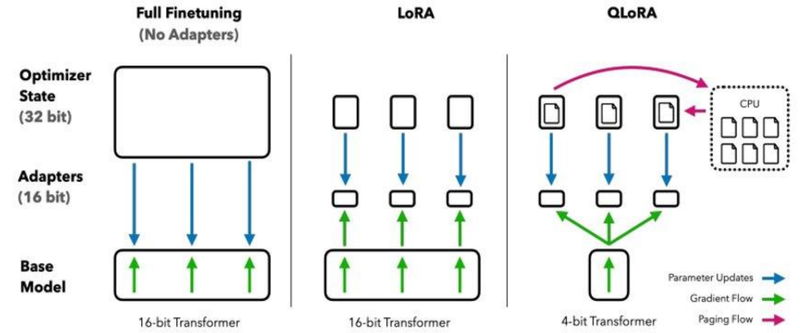


Figure 5: Representation of full parameter fine-tuning (left), LoRA fine-tuning (middle), and QLoRA fine-tuning (right). Full parameter fine-tuning allows for high-precision floating-point numbers (32/16-bits) and LoRA allows for the addition of adapters, which are also high-precision floating-point numbers (16 bits), while QLoRA quantizes weights to lower-precision formats (4 bits) and distributes tasking via CPU paging flow. (Image from https://arxiv.org/pdf/2305.14314)

- QLoRA quantizes weights to lower-precision formats (8/4-bits) when dealing with high-precision floating-point numbers (32/16-bits)
- QLoRA reduces model size and computational hardware requirements required for fine-tuning (training)

What's Next?

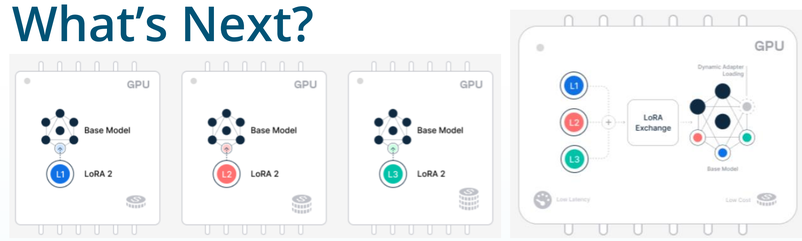


Figure 6: Each LoRA adapter is loaded on their respective GPU along with the base model it was fine-tuned on (left). Combination of the generalized knowledge of the base LLM and multiple domain-specific LoRA "**expert modules**" can be loaded on a single GPU (right). (Image from https://predibase.com/blog/introducing-the-first-purely-serverless-solution-for-fine-tuned-llms)

- Creating fine-tuned "**Expert Modules**"
- Team-Specific | Domain Adaptability | Contextual Understanding
- Utilization of generalized learned knowledge of a base LLM

U.S. DEPARTMENT OF ENERGY
National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

SAND2024-14215C

Sandia National Laboratories