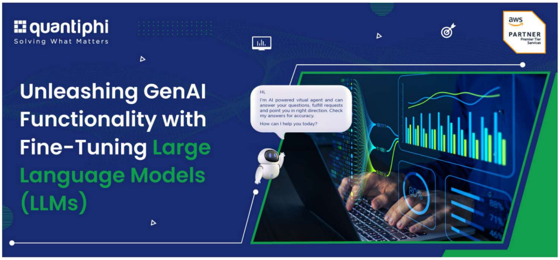
Unleashing GenAI Functionality With Fine-Tuning Large Language Models (LLMs)

Mar 28, 2024

Share

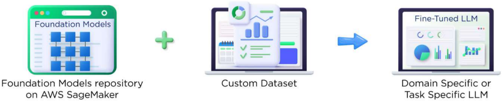

quantiphi
Solving What Matters
aws PARTNER Premier Tier Services
Unleashing GenAI Functionality with Fine-Tuning Large Language Models (LLMs)
Hi,
I'm AI powered vitual agent and can answer your questions, fulfill requests and point you in right direction. Check my answers for accuracy.
How can I help you today?
80%

Embark on an illuminating journey into the world of fine-tuning large language models (LLMs) on AWS, where the power of customization knows no bounds. Fine-tuning isn't just a technique; it's the key to harnessing the full potential of LLMs and adapting them to the unique demands of various industries.

In today's rapidly evolving landscape, the impact of fine-tuned LLMs is profound. In healthcare, it enables the creation of personalized treatment plans that can change lives. In finance, it's the compass guiding precise market predictions, empowering investors to make informed decisions. Legal professionals discover a source of efficiency in their capacity to expedite contract reviews, paving the way for a more productive and promising future. And this is merely the tip of the iceberg in terms of its potential.

In this blog, the Quantiphi team unveils the art of supercharging LLMs to deliver domain-specific outcomes that redefine possibilities. Imagine offering not just personalized diagnostics but also revolutionizing market trend predictions and expediting contract reviews by automating the identification of crucial clauses and potential risks.

Traditional fine-tuning methods have been resource-intensive and time-consuming. However, with Parameter Efficient Fine-tuning (PEFT), an ingenious open-source library from Hugging Face, adapting pre-trained LLMs on AWS for various applications becomes a streamlined journey.


Foundation Models
Foundation Models repository on AWS SageMaker
+
Custom Dataset
Fine-Tuned LLM
Domain Specific or Task Specific LLM

From Good to Great: The Vital Need for Fine-Tuning Large Language Models:

Fine-tuning large language models (LLMs) offers several distinct advantages

> **Domain-Specific Expertise:** By fine-tuning the LLM on specific data related to a particular domain or industry, the model gains a deeper understanding of the domain's jargon, terminology, and context. This leads to precise and pertinent responses customized to the specific domain like fine-tuned model can assist legal professionals by reviewing contracts, identifying potential risks, and suggesting legal language adjustments.

> **Faster Inference:** Fine-tuned models require fewer parameters, resulting in lower memory consumption and quicker inference times. This optimizes speed in real-time applications and large-scale deployments.

> **Enhanced Efficiency:** Fine-tuning allows customers to focus on specific aspects of their model, reducing the need for extensive training from scratch. This results in significant time and cost savings within the development process.

> **Data Privacy and Security:** Fine-tuning in-house data ensures that sensitive information remains within the organization. This reduces the risk of exposure to proprietary or sensitive data to external parties.

> **Adaptability:** Fine-tuned models can be rapidly updated and re-trained with new data to remain current with domain trends. This ensures ongoing performance enhancement.

Challenges:

The fine-tuning of large language models is borne with its own set of challenges. These challenges can be broadly classified into categories such as Data, Technical and Resource related.

> **Training Data:** Fine-tuning of LLM models requires a large amount of diverse and high-quality training data to produce accurate and reliable results. Acquiring and curating such datasets can be time-consuming and resource-intensive.

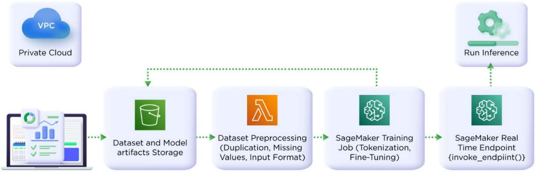
> **Model Complexity**: Generative AI models are complex and resource-intensive during training and deployment, resulting in elevated computational expenses and extended processing times.

> **Ethical challenges:** Fine-tuning of LLM models may inadvertently generate sensitive or personally identifiable information, leading to potential privacy violations. There is a risk of the model generating inaccurate or biased content, which could influence decision-making or propagate misinformation. It can also be misused to create fake documents, signatures, or identities, leading to potentially fraudulent activities. As such, we must dive deep into ethical considerations with commitment to responsible AI deployment.

AWS provides a powerful platform with various tools to help overcome these challenges effectively. It provides easy access to scalable and managed infrastructure, allowing for seamless selection of instance type, number of instances, and distributed training. AWS allows to deploy models easily using endpoints and provides automatic scaling and load balancing. Quantiphi, a SageMaker Service Ready partner, insists on prioritizing diverse and representative data during training, and implements bias detection techniques. It enables us to provide the synthetic data needed to train and enhance LLMs' capabilities, overcoming challenges associated with real data limitations, privacy concerns, and specialized domain requirements.

Process of Fine-Tuning Large Language models:

Fine-tuning a large language model on AWS involves several steps. Below is the detailed process:


VPC
Private Cloud
Run Inference
Dataset and Model artifacts Storage
Dataset Preprocessing (Duplication, Missing Values, Input Format)
SageMaker Training Job (Tokenization, Fine-Tuning)
SageMaker Real Time Endpoint {invoke_endpiint()}

> **Set Up AWS Environment:** Setting up AWS services for fine-tuning a large language model involves multiple steps like creating or configuring aws account, IAM role configuration with necessary permissions for SageMaker to access other AWS services. Next, we can create and set up lambda to create SageMaker training jobs and endpoints.

> **Prepare Dataset:** First, we need to gather and prepare a domain-specific dataset, that is representative of the target task and will store the documents on S3. Next, we can perform data cleaning and preprocessing like duplicates, handling missing values, and normalizing the text to remove any noisy or irrelevant information using containerized AWS lambda. After preprocessing, split the dataset into training, validation, and testing sets and apply the formatting process to pack multiple samples to one sequence for efficient training and then tokenize them using “**AutoTokenizer**” from the “**transformers**” library.

> **Fine-Tune the Model:** Initiating the SageMaker Training Job using the Hugging Face Estimator library to fine-tune the selected large language model using **QLoRA** method from **PEFT** library.
> The **Estimator** streamlines the complete training and deployment process within Amazon SageMaker. It effectively oversees infrastructure management, with SageMaker. It ensures the appropriate Hugging Face container, uploads designated scripts, and seamlessly retrieves data from our S3 bucket. This data is then efficiently integrated into the container at /opt/ml/input/data.
>
> **PEFT (Parameter Efficient Fine-tuning)** an open-source toolkit by Hugging Face designed to streamline the process of adapting pre-trained language models for diverse downstream tasks. This innovative library allows for targeted refinement of specific model parameters, enhancing efficiency without necessitating fine-tuning of the entire model.
>
> QLoRA: Efficient Fine Tuning of Quantized LLMs presents an effective approach for fine-tuning large language models by quantizing a pretrained model to just 4 bits and adding compact “Low-Rank Adapters” that undergo fine-tuning. This innovation empowers the fine-tuning of even massive models containing up to 65 billion parameters using a single GPU. Despite its impressive efficiency, QLoRA performs with full-precision fine-tuning, showcasing exceptional performance on language-related tasks and achieving state-of-the-art results.
>
> We can initiate a training job by using the .fit() method from **Estimator**, where we provide an S3 path to the training script.

> **Deploy the Model:** To deploy the fine-tuned model on SageMaker, We create a HuggingFaceModel model class and define our endpoint configuration including the hf_model_id, image_uri, model_data, instance_type, etc. To get the latest Hugging Face LLM DLC on Amazon SageMaker, we can utilize the get_huggingface_llm_image_uri function available through the SageMaker SDK and model_data is the S3 path of fine-tuned model artifacts. Based on the model size, we can select the instance_type (ml.g5 series instances), to create the SageMaker real time endpoint. Once the endpoint is deployed and it’s in “In-Service” state, we can execute inferences using the predict method.

Conclusion

In conclusion, the potential of GenAI functionality is truly revolutionary. By harnessing the capabilities of cutting-edge language models, professionals across various industries can achieve unparalleled domain mastery and ignite innovation. This transformation will reshape how we approach fields as diverse as healthcare and marketing, expanding horizons and pushing the boundaries of what's possible.

Quantiphi is at the forefront of AI innovation, helping businesses across all industries to unlock the power of cutting-edge AI and Machine Learning technologies. As AWS's established ML/AI partner of the year and wealth of experience and expertise in AI and cloud, Quantiphi enables you to harness the full potential of Generative AI on AWS to drive growth and success in your transformative journey.

WRITTEN BY
Stephanie Pace

FOUND THIS USEFUL? SHARE WITH