Building Next-Generation LLM Applications with Private Models

Using the Reasoning Workbench to Better Select Private LLM Models

Josh Patterson

Section 1: Introduction

At AWS Re:Invent 2023, the Generative AI landscape witnessed a wave of innovations, and among the standout announcements, AWS SageMaker and AWS Bedrock emerged as go-to platforms for deploying cutting-edge Generative AI applications. As we step into 2024, the spotlight turns to the compelling investment in Large Language Models (LLMs) for reasoning applications. The potential impact on enterprise productivity, market share, and profitability is substantial, prompting a closer examination of both the opportunities and challenges that come with embracing this transformative technology.

Section 2: Trends

In the fast-paced world of Generative AI, AWS Re:Invent 2023 set the stage for the ascendancy of AWS SageMaker and AWS Bedrock. These platforms are not just tools; they are enablers of a new era, providing a robust foundation for deploying applications powered by Large Language Models (LLMs). The applications of LLMs are diverse, from generating human-like text and answering questions to assisting in language translation, code writing, summarizing articles, and crafting conversational agents. The game-changer here is the integration of LLMs with enterprise data, employing the RAG design pattern on AWS to seamlessly connect private knowledge repositories with the unparalleled reasoning prowess of LLMs. This dynamic integration opens up avenues for innovation, reshaping how enterprises harness the potential of Generative AI.

Section 3: What to Look Out for When Building LLM Applications

As we embark on the journey of building LLM applications, a crucial perspective emerges – the interplay between "knowledge" and "reasoning." In the context of enterprise applications, the emphasis is on augmenting prompts with private data from knowledge repositories like data warehouses and S3. However, a pivotal consideration arises when choosing between using someone else's model, like the powerful GPT3 or GPT4 from openAI, and opting for smaller private models. The trade-off involves efficiency versus horsepower in reasoning tasks. Striking the right balance is essential for building efficient, cost-effective AI applications.

Section 3.1: Challenges in Choosing a LLM Model

While openAI's models bring immense power, deploying them as private LLM model instances can be challenging. The alternative of smaller private models introduces efficiency but may lack the robust reasoning capabilities required for certain tasks. The key lies in evaluating different models for specific tasks with enterprise data. Enter the Reasoning Workbench, an open-source tool designed by Patterson Consulting, providing a streamlined approach to compare models efficiently.

Section 3.2: The Reasoning Workbench

Patterson Consulting's Reasoning Workbench emerges as a beacon in the quest for choosing the right private LLM model. This open-source tool simplifies the process of generating scores to compare different models. With a focus on efficiency, the Reasoning Workbench empowers teams to swiftly identify the most suitable model for their specific AI application, ensuring a judicious balance between performance and operational cost.

Section 3.3: Benefits of Evaluating Models with the Reasoning Workbench

The ability to expeditiously select the best model for a use case equips teams to deploy new AI applications confidently and decisively. Beyond agility, the cost-efficiency of better models, with fewer tokens and lower cloud resource requirements, underscores the economic advantages. Moreover, the streamlined model evaluation process reduces the need for extensive prompt engineering, culminating in a faster and more efficient user experience.

Section 4: Using the Reasoning Workbench on AWS

Transitioning from theory to practice, integrating the Reasoning Workbench into AWS SageMaker Studio Lab brings the power of efficient model evaluation to the fingertips of developers and data scientists.

Section 4.1: Writing a Configuration File

Configuring the models.json file and prompts.json file becomes the initial step, where each entry corresponds to a model hosted on HuggingFace, providing a structured approach to harnessing the capabilities of different models.

Section 4.2: Generating the Model Report

Running the notebook triggers the generation of a comprehensive model report, including JSON files with performance metrics, radar plot images showcasing general and use case-specific performance, and a nuanced understanding of how models fare across various sub-tasks.

Section 5: Interpreting the Report

Diving into the generated report, the general performance radar plot offers a relative perspective on different private models, encompassing Language Understanding, World Knowledge, Math, and Correctness. The innovative aspect lies in the ability to customize prompts for evaluating models, enabling a nuanced comparison tailored to the organization's unique data and domain.

Section 5.1: Implications of Model Metrics

Metrics like Average Prompt Inference Time and Model Parameter Size become pivotal in shaping decisions on GPU selection and operational costs. Balancing prompt analysis speed with GPU memory requirements guides the selection of an optimal GPU, aligning technical considerations with financial prudence.

Section 6: Next Steps

Armed with a baseline comparison from the Reasoning Workbench, the journey into the realm of Generative AI unfolds with a myriad of next steps.

Section 6.1: Deploy to AWS SageMaker

For those ready to experiment further, deploying specific models on AWS Bedrock becomes a tangible next step, facilitated by the EZSMDEPLOY Python SDK.

Section 6.2: Help Designing Custom Prompts

Exploring custom model performance on specific use cases finds support in Patterson Consulting, offering expertise in custom prompt development tailored to unique organizational requirements.

Section 6.3: Designing and Deploying Custom Agents on AWS

The path to building LLM applications on AWS takes shape with Patterson Consulting's assistance in designing and deploying custom agents, ensuring a seamless integration with existing teams and workflows.

Section 6.4: Fine-Tune a Model

Further refinement becomes possible with Patterson Consulting's expertise in fine-tuning specific models for specific use cases, adding a layer of customization to elevate AI applications to new heights.

In conclusion, the journey into the realm of Large Language Models for reasoning applications is a transformative expedition, and with the right tools and strategic considerations, enterprises can navigate this landscape with confidence, unlocking a realm of possibilities for innovation and efficiency.