**Notes on LLM Execution and Inference Engines**

1. Understanding LLM Execution

To run a Large Language Model (LLM), you need both:

1. The Model – The actual neural network (e.g., CodeLlama, GPT-4, LLaMA, Falcon).
2. An Inference Engine / Runtime – Software that loads and runs the model efficiently (e.g., Ollama, Hugging Face, vLLM).

2. Inference Engines & Deployment Options

Popular Inference Engines:

* Ollama – Local runtime for running LLMs efficiently.
* Hugging Face TGI – Hosted & self-managed deployment for open-source models.
* vLLM – Optimized runtime for fast text generation using FlashAttention.
* DeepSpeed-Inference – Microsoft’s high-speed inference engine for transformers.
* TensorRT-LLM – NVIDIA’s optimized runtime for inference on GPUs.

Azure-Based Alternatives:

* Ollama (Local Inference) → Azure Machine Learning + Custom VM/AKS
* Hugging Face API → Azure OpenAI Service
* Self-Hosting LLMs → Azure Kubernetes Service (AKS) or Azure Functions

3. CodeLlama & Hugging Face

* Is CodeLlama GPT-based? → No, it is based on LLaMA, a transformer model by Meta.
* Is Hugging Face like Ollama? → No, Hugging Face provides model hosting & APIs, while Ollama is a local runtime.
* Does Hugging Face host CodeLlama? → Yes, it hosts multiple LLMs including CodeLlama.

4. How LLM Loading Works Technically

When an LLM is loaded into an inference engine:

* Model Architecture is Initialized → Loads the transformer structure.
* Pretrained Weights are Loaded → Retrieves trained parameters into RAM.
* Memory is Allocated → Space for computations is allocated on CPU/GPU.
* Model is Optimized → Techniques like quantization, batching, and FlashAttention are applied.
* Tokenizer is Loaded → Converts input text into numerical token sequences.
* Model Execution Begins → Text is fed into the model, and predictions are generated.

5. Running CodeLlama on Azure

To build a CodeLlama-powered query system on Azure:

* Inference Engine → Azure Machine Learning or Azure OpenAI Service.
* Model Storage → Azure Blob Storage or Hugging Face Hub.
* Execution Platform → Azure Kubernetes Service (AKS) or Virtual Machines (VMs).
* API Deployment → FastAPI inside a Docker container on Azure.