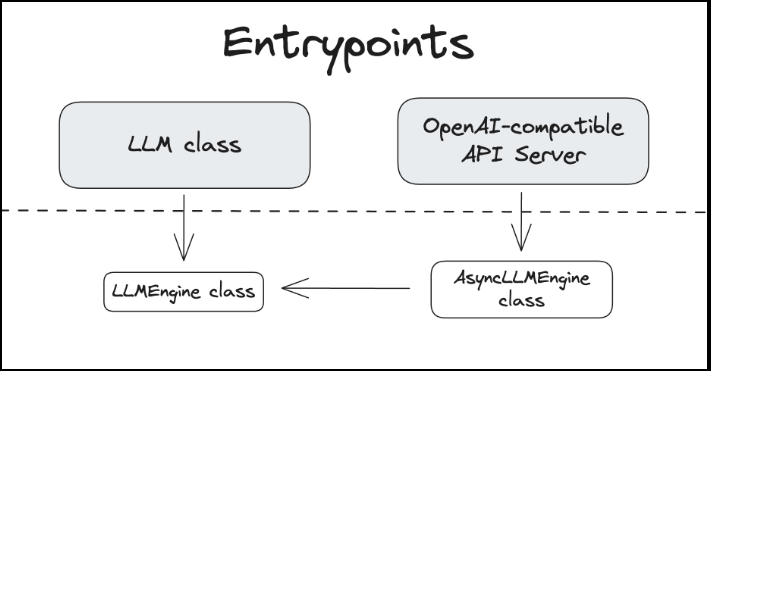
**1. Title Slide**

* Title: *"vLLM: Fast & Cost-Efficient LLM Serving"*
* Subtitle: *How PagedAttention and Continuous Batching Optimize LLMs*

VLLM Architecture overview:

[LLM Icon] → [vLLM Icon] → [$$$]│



<https://blog.runpod.io/introduction-to-vllm-and-how-to-run-vllm-on-runpod-serverless/>

<https://medium.com/cj-express-tech-tildi/how-does-vllm-optimize-the-llm-serving-system-d3713009fb73>

from transformers import AutoTokenizer, AutoModelForCausalLM

import torch

# Load tokenizer and model

model\_name = "facebook/opt-125m"

tokenizer = AutoTokenizer.from\_pretrained(model\_name)

model = AutoModelForCausalLM.from\_pretrained(model\_name)

# Move model to CPU or GPU

device = torch.device("cuda" if torch.cuda.is\_available() else "cpu")

model.to(device)

# Input prompt

prompt = "In the future, artificial intelligence will"

# Tokenize input

input\_ids = tokenizer(prompt, return\_tensors="pt").input\_ids.to(device)

# Generate output

output\_ids = model.generate(input\_ids, max\_length=50, do\_sample=True, temperature=0.7)

# Decode and print

output\_text = tokenizer.decode(output\_ids[0], skip\_special\_tokens=True)

print(output\_text)

**Why is serving LLM so challenging?**

**Computational Resources**

Due to the fact that LLM has numerous parameters to perform a prediction, which could start with the 7B parameter and then go up to 321B, deploying this model may require an intensive resource and a lot of optimization rather than using a traditional method to deploy a machine learning model.

**Latency**

When a sentence or token is complicated, the process takes several minutes to compute a result for the client, which may cause an issue on a large scale or in real-world business. For instance, a company may apply LLM with a product Q&A chatbot, which has a slow response to each question, which could cause frustration for the user. Therefore, applying some method to reduce the latency would be a good practice.

**Cost**

In a large-scale system or with multiple LLMs in the system, which would consume a lot of budget for the application since LLMs use large resources to process, as a MLE, finding a way to utilize a resource would bring a financial benefit to the system. For instance, lower the cost per request.

**What is vLLM?**