**System Design Tasks:**

**§ Design system architecture to deploy this ML Model in produc1on**

To design the system architecture which is highly scalable we can do

We can use either AWS lambda or containerize the model and then deploy it with Kubernetes.

And before containerizing the model we can also create API endpoints of the model to handle incoming requests from clients and route them to the appropriate endpoint for model inference

For the real time system, we can build the pipelines using tools like airflow for real time data cleaning and feature engineering.

For monitoring and logging, we can use AWS CloudWatch for monitoring AWS services and resources and for model we can use mlflow for model tracking and model versioning

For monitoring and visualization we can create dashboards that display real-time metrics and performance indicators.

**§ How do you perform canary build?**

This can be achieved by deploying the new model container to a separate environment or a fraction of the container servers and gradually increasing the traffic to it while monitoring key metrics. If the new version performs well, it can be promoted to full deployment otherwise, it can be rolled back.

**§ What should be the strategy for ML Model Monitoring?**

We can implement monitoring solutions to track various performance metrics of the deployed ML model, including accuracy, latency and resource utilization.

We can also set up alerts and notifications to trigger when performance metrics deviate from predefined thresholds, indicating potential issues and reduction in model performance

And Always ensure to continuously monitor data drift and model drift to ensure that the deployed model remains effective and accurate over time.

**§ How do you perform load and stress tes1ng?**

We can conduct load and stress testing to evaluate the scalability and robustness of the system under various levels of load and traffic.

Or simulate high traffic scenarios and measure the response times, error rates, and resource utilization of the system to identify potential areas for optimization.

**§ How do you track, monitor and audit ML training?**

I usually use experiment tracking platforms like MLflow or TensorBoard to log and monitor the performance of different model versions during training.

**§ Design framework for con1nuous delivery and automa1on of machine learning tasks.**

Firstly we can set up a CI/CD pipeline to automate the deployment of new model versions to production.

Secondly we can automate model retraining and deployment processes using tools like Jenkins, GitLab CI/CD or AWS CodePipeline or build the airflow pipeline