**One-Day Workshop Agenda: Model Deployment and Explainability**

**Conceptual Foundations**

1. **Utilization of Self-Attention in Transformer Encoders and Decoders**
2. **Scaled Dot-Product & Multi-Head Attention**
3. **Positional Encoding and Layer Normalization**

**Explainability & Interpretation**

1. **Model Explainability & Transformer Visualization**
2. **Attention Visualization using Hugging Face**

**Optimization & Deployment**

1. **Deployment Considerations**
2. **Model Optimization (Pruning, Quantization, Knowledge Distillation)**
3. **Inference Efficiency**
4. **Cloud Deployment Options**
5. **Evaluation Metrics – Accuracy, Latency, Throughput**