Deploying Small Language Modules (SLMs) on Edge Devices

Objective:

The primary objective of this internship was to carry out various metrics of Small Language Modules (SLMs) such as MobileBERT, DistilBERT, and TinyBERT on edge devices to evaluate their performance in terms of accuracy, speed, and resource utilization. This project aimed to identify the most suitable model for edge deployment, balancing efficiency and accuracy.

Tasks Undertaken:

1. **Research and Model Selection:**
   Conducted a comprehensive analysis of MobileBERT, DistilBERT and TinyBERT to understand their suitability for edge devices.

2. **Understanding the Models:**
   Successfully understand the pros and limitations of the models. Different model has different speed, accuracy and memory usage.

3. **Metrics Calculation:**
   Designed and executed experiments to measure various performance metrics, including:
   - Model accuracy
   - Inference time (speed)
   - Memory usage
   - Power consumption

4. **Results Analysis:**
   Compared the performance of MobileBERT, DistilBERT, and TinyBERT on key metrics.

Key Findings:

- **Accuracy:** DistilBERT outperformed other models in terms of accuracy, making it ideal for applications where precision is critical.
- **Speed:** TinyBERT was the fastest model during training and inference, making it highly suitable for resource-constrained scenarios.
- **Resource Utilization:** MobileBERT demonstrated a balance between accuracy and resource requirements, serving as a versatile option for edge deployments.

Outcomes and Contributions:

- Successfully understand the basics of SLMs.
- Identified the trade-offs between accuracy and efficiency across different models, providing actionable insights for future projects.
- Developed scripts and tools for automating the measurement of metrics on edge devices.

Conclusion:

This internship allowed me to enhance my understanding of deploying machine learning models on edge devices and contributed to the advancement of efficient language model usage in constrained environments.