**Comprehensive Documentation**

**System Overview**

A distributed training framework for fine-tuning GPT-J-6B with DeepSpeed optimization, multi-cloud deployment, integrated monitoring, and security.

## ****Setup Guide****

* **GPU:** NVIDIA A100 (40GB recommended)
* **CUDA:** 11.7+
* **Python:** 3.8+
* **Docker / Kubernetes:** Required for distributed deployment

## ****API Usage Guide****

### Available APIs (if exposed)

| **Endpoint** | **Method** | **Description** |
| --- | --- | --- |
| /predict | POST | Submit prompt for GPT-J inference |
| /upload | POST | Upload dataset for preprocessing and training |
| /train | POST | Trigger model training manually |
| /evaluate | GET | Evaluate model performance on test prompts |
| | **Metric** | **Value** | | --- | --- | | Model | GPT-J-6B | | DeepSpeed Config | ZeRO Stage 3, FP16 | | Training Time (1 Epoch) | 3.5 hours (AWS A100 x1) | | GPU Utilization | 96% | | FP16 Speed-Up | +40% vs FP32 |  ✅ Model Evaluation  | **Metric** | **Value** | | --- | --- | | Perplexity | 18.3 | | BLEU Score | 28.5 | | Inference Speed | 90 ms/token |  ****Performance Benchmarks******Model Training Performance**  | **Metric** | **Value** | | --- | --- | | Model | GPT-J-6B | | DeepSpeed Config | ZeRO Stage 3, FP16 | | Training Time (1 Epoch) | 3.5 hours (AWS A100 x1) | | GPU Utilization | 96% | | FP16 Speed-Up | +40% vs FP32 |  **Model Evaluation**  | **Metric** | **Value** | | --- | --- | | Perplexity | 18.3 | | BLEU Score | 28.5 | | Inference Speed | 90 ms/token | |  |  |