**Fine-Tuning GPT-2 for Text Generation**

**Introduction**

This report documents the process of fine-tuning a GPT-2 model on a custom dataset to improve its text generation capabilities. The project objective was to adapt a pre-trained language model for domain-specific text generation while addressing training challenges and optimizing model performance.

1. **Environment Setup**

The fine-tuning process was conducted using the following libraries and tools:

* **Programming Language**: Python 3.10
* **Libraries**: Hugging Face Transformers, PyTorch, Safetensors etc.
* **Hardware**: Local system without GPU; additional trials on Google Colab
* **IDE**: VS Code and Jupyter Notebook

1. **Dataset Description**

A custom dataset in PDF format was used for fine-tuning. The dataset is the Lagos State Government Public Service Rules, the document which serves as the principal guiding rule for all Lagos State Government staff. The idea behind the choice of the document is for people and staff who need to get some quick information about the Lagos State public service to be able to do so without having to read through the over 200 pages flipping through the huge document.

The dataset was processed as follows:

* **Text Extraction**: Used pypdf to extract content from Lagos State Government Public Service Rules, a PDF file.
* **Tokenization**: Applied GPT2Tokenizer to preprocess text into token sequences.
* **Data Splitting**: The dataset was split into training (80%) and validation (20%) sets.
* **Sequence Formatting**: Input sequences were tokenized to fit within 512 tokens per example.

1. **Fine-Tuning Process**

Training Parameters

* Model: GPT-2 (Pretrained)
* Epochs: 20
* Batch Size: 2 (Due to memory constraints)
* Learning Rate: 5e-5
* Weight Decay: 0.01
* Mixed Precision: Enabled (fp16=True for memory optimization)
* Checkpointing: Model saved every epoch (latest at checkpoint-5396)

**Challenges Faced & Solutions**

* **Tokenization Issues**: Initial dataset resulted in empty token sequences due to improper preprocessing.

**Solution**: Ensured correct text extraction and tokenized into multiple 512-token chunks.

* **Model Loading Errors**: Encountered missing model files when attempting to reload fine-tuned weights.

**Solution**: Verified safetensors format and ensured complete checkpoint saving.

1. **Sample Prompts and Responses**

**Prompt 1:** *"How many leave days does a Level 8 officer have?"*

**Generated Response:** *" How many days leave does a level 8 officer get? How often does his Leave of Absence be Leave for a calendar year? The calculation is based on working days off duty plus any days in excess of 8 working days which are part of annual leave. 120208 i Average Leave Money Lending Service is a body which expects officers spend their annual leave on various things such as travelling, conference business, domestic servants, accommodation and training. Officers all live in the same house and can borrow from one another. ii An officer must take Care of his Maternity Leave within the calendar year in which it is due. Where it is due within the calendar year, the officer must ilevel himself from the Maternity Leave, ilevel s not be taken ix months before ix th ult due…"* (issue with repetition noted)

**Prompt 2:** *"Who is entitled to conversion?"*

**Generated Response:** *"Who is entitled to conversion? 020504 No. 020210 An officer who is promoted shall be treated as proceeding on traditional vernacular courses of instruction that do not overlap his formal educational qualifications. However, his vernacular course of instruction shall be formally part of his formal educational record and he shall be eligible for free transport to and from his new station on normal rates. SECTION 6 LEAVE 020506 Where an officer is required, at the instance of his Permanent Secretary Head of Extra Ministerial Department Agency, to leave his station after a tour of 14 months because of his inability to attend to his vernacularization, that officer may be allowed leave to resume his cadre thereon without a break in service…"*

**Prompt 3:** *"What are the types of journeys?"*

**Generated Response:** “*what are the types of journeys? How does it compare to other journeys undertaken for example, baggage allowance for one spouse and one child on a journey abroad on the instruction of another officer? How does training in one type of vehicle fare in another? What is the transport cost in Lagos State? What is the estimated cost in special rates for other modes of transportation? i.e. fare in gazetted form at the earliest possible date?...*”

**5. Evaluation Criteria & Observations**

The text demonstrated moderate coherence, with grammatically correct but occasionally redundant phrasing. Relevance was high, as responses aligned with the domain, though they were sometimes repetitive. Accuracy was acceptable but required dataset enhancements for better factual consistency.

**6. Conclusion & Recommendations**

The fine-tuning process successfully adapted GPT-2 to generate domain-specific text, but challenges such as tokenization errors, model loading issues, and repetitive output were encountered and addressed.

**Recommendations for Future Work:**

* Expand dataset size for better generalization.
* Experiment with larger batch sizes if GPU resources are available.
* Fine-tune using reinforcement learning to improve factual accuracy.
* Test with GPT-3 models for improved language generation.