Technical Documentation for Fine-Tuned BERT Model

1. Introduction

- **Purpose:** This document provides a comprehensive overview of the fine-tuned BERT (Bidirectional Encoder Representations from Transformers) model developed to enhance the IHEC Student Chatbot's capability to understand and respond to student inquiries in French.
- **Scope:** The model is integrated into the IHEC website to address frequently asked questions, delivering accurate and timely information to students.

2. System Overview

- **Architecture:** The model is based on the BERT architecture, specifically the multilingual version (bert-base-multilingual-cased). Fine-tuning was performed to adapt the model to the specific dataset of student inquiries.
- **Components:**
  - **Tokenizer:** Utilizes the BertTokenizer from the Hugging Face Transformers library to convert input text into tokens compatible with the BERT model.
  - **Neural Network Layers: Consists** of multiple transformer layers that process the input tokens, capturing contextual relationships within the text.
  - **Output Layer:** A classification layer that maps the processed information to predefined categories of answers, enabling the model to select the most appropriate response.

3. Data Management

- **Dataset Description:** The dataset comprises pairs of questions and corresponding answers related to student inquiries at IHEC. Each entry includes an ID, category, question, and answer.
- **Data Storage :**The dataset is stored in a JSON file format, ensuring compatibility with the model's input requirements.

4. Data Preprocessing

- **Text Cleaning:**Text data is preprocessed by converting to lowercase and removing special characters to standardize the input and reduce noise.
- **Tokenization:**The BertTokenizer is employed to tokenize the questions, converting them into a format suitable for the BERT model.
- **Dataset Creation:**A custom FAQDataset class is implemented to handle tokenized inputs and labels, facilitating efficient data loading during training.

5. Model Training

- **Training Process:**

    - **Data Splitting:**The dataset is divided into training and validation sets to evaluate the model's performance on unseen data.
    - **Model Initialization:**The BertForSequenceClassification model is initialized with the pre-trained multilingual BERT model and configured for the specific classification task.
    - **Optimizer and Scheduler:**The AdamW optimizer is used with a learning rate scheduler to adjust the learning rate during training, promoting efficient convergence.
    - **Training Loop:**The model is trained over multiple epochs, with loss computation and backpropagation performed in each iteration to update model parameters.
- **Hyperparameters:**
    - **Learning Rate:**2e-5
    - **Batch Size:**16
    - **Epochs:**3
- **Performance Metrics:** Accuracy is used to evaluate the model's performance during training, with validation accuracy monitored to prevent overfitting.

6. Model Evaluation

- **Validation:**The model's performance is assessed using a validation set to ensure generalization to unseen data.
- **Results:** The model achieves satisfactory accuracy, indicating effective adaptation to the dataset.

7. Deployment

- **Integration:**The fine-tuned model is integrated into the IHEC website via an API, allowing real-time responses to student inquiries.
- **Environment:**The deployment environment includes necessary software dependencies and hardware specifications to support the model's operation.

8. Maintenance and Updates

- **Monitoring:**Regular monitoring is conducted to track the model's performance and user interactions, identifying areas for improvement.
- **Retraining:**The model is periodically retrained with new data to maintain and improve performance, ensuring it remains up-to-date with evolving student inquiries.
- **Version Control:**Version control practices are employed to manage updates and ensure reproducibility, facilitating collaboration and tracking changes over time.

9. Security and Compliance

- **Data Privacy:**The model complies with CCK regulations, ensuring that no personal user data is stored or processed on a different server, thereby protecting user privacy.
- **Access Control:**Access to the model and its data is restricted to authorized personnel only, with role-based access controls implemented to safeguard sensitive information.

10. Troubleshooting and Support

- **Common Issues:**Potential issues include data inconsistencies and integration challenges, with predefined solutions available to address these problems.

- **Support Contacts:**For assistance, contact the technical support team via the provided communication channels, ensuring prompt resolution of issues.