# Step-by-Step Approach to Solving the CpG Site Prediction Problem

## Step 1: Research and Understanding the Project

Before implementing the solution, I conducted research to understand the core concepts of the project:

* **Understanding CpG Sites:** CpG sites are regions of DNA where a cytosine nucleotide is followed by a guanine nucleotide. These sites play a crucial role in gene regulation and epigenetics.
* **Application of CpG Sites:** CpG methylation is an important factor in diseases like cancer and is widely used in biological research and diagnostics.

## Step 2: Choosing LSTM for the Task

Once I understood the biological significance of CpG sites, compare to RNN , Lstm is good for the sequence task . RNN Have Vanishing and Exploding Gradients problem issue.

* LSTMs are effective for sequential data like DNA sequences.
* They can capture long-range dependencies better than traditional neural networks.
* CpG site prediction involves analyzing patterns in DNA sequences, making LSTM a suitable choice.

## Step 3: Handling Variable-Length DNA Sequences

One challenge in DNA sequence prediction is handling variable-length sequences. To address this:

* I used **padding** to ensure all sequences are of the same length before passing them into the LSTM model.
* I explored **embedding techniques** to represent DNA sequences in a numerical format.
* I applied **batch processing strategies** to efficiently train the model on sequences of different lengths without loss of information.

## Step 4: Understanding and Completing the Code

* I carefully studied the existing code to understand the architecture and data pipeline.
* I identified missing parts and completed them accordingly.
* I debugged errors using Stack Overflow and ChatGPT to resolve issues related to tensor shape mismatches, model architecture, and data preprocessing.

## Step 5: Selecting Optimal Hyperparameters

After debugging, I focused on selecting appropriate hyperparameters for training the LSTM model. I experimented with different values and optimized the following:

* **Learning Rate:** To ensure the model converges at an optimal pace.
* **Weight Decay:** To prevent overfitting.
* **Number of LSTM Layers:** To balance complexity and performance.

## Step 6: Training the Model

After setting the right hyperparameters, I trained the model using:

* **Loss Function:** Mean Squared Error (MSE) since the task involves numerical predictions.
* **Optimizer:** Adam optimizer was chosen for its adaptive learning rate and efficiency in training deep networks.
* **Training Parameters Used:**

## Step 7: Model Evaluation

Once training was complete, I evaluated the model using the Mean Absolute Error (MAE) metric because:

* MAE provides a more interpretable error measurement in regression tasks.
* It is less sensitive to large errors compared to MSE.
* It gives a clear idea of how much the predicted CpG site values deviate from the actual values.

## Step 8: Validating Model Performance

* I tested the model on unseen data to verify its predictions.
* I checked whether it was correctly identifying CpG patterns.
* If the performance was not satisfactory, I revisited hyperparameter tuning and debugging.

## Step 9: Deploying the Model with Streamlit UI

After successfully training the model, I followed these steps to deploy it:

* **Saved the trained model** so it could be used for inference.
* **Developed a Streamlit UI** where users can input DNA sequences and get predictions for CpG sites.
* **Deployed the model on Streamlit Cloud** to allow easy access via a web-based interface.

## Conclusion

Following this structured approach, I successfully implemented the CpG site prediction model using LSTM. Handling variable-length DNA sequences, step-by-step debugging, and hyperparameter optimization significantly improved the model’s accuracy. Finally, integrating a web UI using.