Report on Named Entity Recognition Model

Current Issues:

Data Quality and Diversity:

- The current dataset creation process involves manual extraction of sentences and labels from ChatGPT, leading to a limited and potentially biased dataset.
- Recommendation: Accessing OpenAI directly through a .py script would provide a more diverse and representative dataset. This can be achieved by generating sentences programmatically using the ChatGPT API, ensuring a broader spectrum of language patterns and entity mentions.

Tokenization Process:

- The current tokenization approach may not adequately capture the nuances of sentence structure, leading to suboptimal model performance.
- Recommendation: Revisit the tokenization process to ensure that the model receives more meaningful input. Consider using word-level tokenization or exploring alternatives that better represent the underlying semantics of the sentences.

Model Choice:

- Training the model on distilbert-base-uncased may not be the optimal choice for Named Entity Recognition tasks, especially when specialized models pre-trained for NER exist.
- Recommendation: Explore and fine-tune models specifically designed for NER tasks, such as BERT-based pre-trained models fine-tuned on NER datasets. This change can significantly improve the model's ability to recognize mountain names accurately.

Proposed Improvements:

Data Collection:

- Utilize a script to interact with the ChatGPT API for automatic dataset generation.
- Ensure the dataset includes a diverse set of sentences containing mountain names in various contexts.

Tokenization Enhancement:

- Experiment with different tokenization techniques, such as word-level tokenization, to better represent the linguistic structure of sentences.
- Explore specialized tokenization libraries designed for NER tasks.

Model Fine-Tuning:

- Choose a pre-trained model that excels in NER tasks, such as BERT or other state-of-the-art models.
- Fine-tune the selected model on the generated dataset for optimal performance in mountain name recognition.

Evaluation Metrics:

- Implement comprehensive evaluation metrics to assess the model's performance accurately.
- Metrics such as precision, recall, and F1 score can provide insights into the model's ability to correctly identify mountain names.

Conclusion:

By addressing the mentioned issues and implementing the proposed improvements, the Named Entity Recognition model for mountain names can achieve better accuracy and generalization. It is essential to prioritize data quality, choose a suitable pre-trained model, and fine-tune it on a well-curated dataset for optimal performance.