Task 4

Rationale

1. **Pre-trained BERT Model**:
    - **Lower Learning Rate (e.g., 1e-5)**: In case of training the BERT model, the layers are already well-trained on a large corpus of text. A lower learning rate prevents high hallucinations or forgetting by making small adjustments to these weights, thus preserving the valuable features it has already learned which are useful across both tasks.
2. **Task-Specific Heads**:
    - **Higher Learning Rate (e.g., 3e-4)**: Since the heads are task-specific and are starting from scratch (not pre-trained), they need a higher learning rate to speed up their convergence during training. This allows these layers to learn task-specific nuances more effectively.

Potential Benefits

- **Customized Learning**: Different layers can learn at rates optimal for their specific situation — pre-trained layers can gently adjust, while new layers can rapidly learn.
- **Prevent Overfitting in Lower Layers**: By using a lower learning rate for the pre-trained parts of the network, we can reduce the risk of overfitting these layers to the new task-specific data.
- **Multi-Task Efficiency**: In a multi-task setting, where different heads may serve different types of tasks, having distinct learning rates ensures that each head can optimize its learning speed without affecting the learning of others, which is crucial when these tasks have different complexities and data characteristics.

Summary

In Task 4, I implemented layer-wise learning rates for the MultiTaskSentenceTransformer to optimize training across different components of the model. Recognizing that the pre-trained BERT backbone already possesses robust linguistic capabilities, I assigned it a lower learning rate of 1e-5. This approach helps mitigate the risk of catastrophic forgetting by making only minor adjustments to the well-tuned pre-existing features, essential for maintaining performance across tasks. Conversely, the task-specific heads, which are crucial for sentence classification and sentiment analysis and start from an uninitialized state, were assigned a higher learning rate of 3e-4. This facilitates quicker convergence, enabling these heads to adapt rapidly to the nuances of their respective tasks.The method allows for customized learning where each layer learns at an optimal rate for its function, prevents overfitting in more stable, pre-trained layers, and enhances multi-task efficiency by allowing each task to progress without hindrance from others. This method ensures that each part of the model is trained effectively, maintaining balance between leveraging existing knowledge and acquiring new, task-specific information.