Performance Metrics

S. No	Model Type	Training Loss (Classification)	Training Loss (Distillation)	Training Loss (Cosine)	Test Set Performance (Accuracy)
1	Odd Layer	0.5431	0.2795	0.0897	0.9059
2	Even Layer	0.5475	0.2767	0.0910	0.8949
3	LoRA	0.4476	0 (Not used)	0 (Not used)	0.8702

Performance Discussion

1. Odd layers distillation achieves the highest test set accuracy performance among the three, which might be attributed to preservation of more useful knowledge (linguistic or semantic features) during distillation.
2. LoRA has the lowest test set accuracy among the three which might suggest overfitting. It might also suggest that since no distillation or cosine alignment is used, the model doesn't benefit from teacher guidance.

Challenges Faced

1. In the distillation training, both odd and even layers, the averaging of the classification loss, distillation loss and cosine loss with equal weightage might lead to undermining the actual training performance.
2. In the LoRA configuration implementation, the adapters are applied without specifying which modules (query, key, value, output) are adapted.

Suggestions

1. Instead of giving equal weightage to each distillation loss type, learnable weights should ideally be used for the loss components which are updated during training to reflect proper weighed importance of each term. Finally, a weighted average should be taken.
2. In the LoRA configuration implementation, better control could be implemented by specifying the models that would be adapted. This could avoid adapting embeddings, classification heads, or less impactful layers.