**Why is this an important area of study? [0.5 points]**

Transformer is used in almost all areas of deep learning nowadays. But it has many performance bottlenecks, and this prevents large models from being deployed on small devices. Therefore, optimization techniques for transformers are extremely important for production.

**Describe two different techniques/approaches discussed [0.5 points]**

1. Besides what we’ve learned in previous classes, we can avoid computing the full attention matrix by learning sparse attention patterns. This can reduce inference time since attention has a time complexity of O(N^2).
2. We can downsample immediate representations by removing parts of the sequence. This reduces the sequence length by a lot and therefore reduces inference time.

**Discuss relative strengths [0.5 points] and weaknesses [0.5 points] of the two techniques described above. [1 point total]**

Technique 1 is effective on many occasions because the attention scores are usually very sparse. But it still cannot handle very long sequences, this limitation could be especially obvious for autoregressive models.

Technique 2 can potentially allow the model to handle long sequences because we don’t need to process the entire sequence. However, the method for determining which tokens are not necessary can affect the performance.