Problem 1

- Analyzing the final 10 layers Head_0 and Head_3 (the first and fourth heads) usually more dominant and thus their insights are more critical.
- Via the ablation study each heads seems to be contributing insights equally at 0.65% accuracy.
- Head_1 and Head_2 have more off diagonal activations thus they are able to carry more information across the string.
- At 4 heads we might be able to maintain decent accuracy if we remove the heads that have carry information but this is not guaranteed thus keeping all 4 heads is important.

Problem 2

- For shorter sequences the learned positional encoding performs better but as size of the sequence increases the no position encoding and sinusoidal encoding perform better. While sinusoidal encoding performs slightly better the difference negligible.
- The sinusoidal encoding scales better than learned encoding as it maintains an unbiased encoding regardless of the sequence length where as the learned encoding is simply learning how the training data relates to itself thus its will tend to overtrain over the training data. This usually means that we are just learning a lookup table over the training data thus its effect over new data will be at best be none or at worst be adversarial.
- At length 32 the accuracy of the learned encoding is better but this changes at length 64 where no positional encoding and sinusoidal encoding become more accurate and this trend continues for all the larger lengths.