Task 3: Training considerations

1. If the entire network should be frozen.
    a. Implications:
        i. no parameters of the model are updated during training. The model will perform inference using only the pre-trained weights.
    b. Advantages:
        i. This model will perform faster and have less computational costs since no training occurs. This is advantageous for simple tasks or similar tasks to which the transformer was previously trained on.
    c. Disadvantages:
        i. This model cannot adapt to specific tasks
    d. Key Decisions and insights
        i. This uses minimal computational resources; however the power of the transformer model will be severely limited due to its lack of adaptability. The model should only be trained this way if the task is very simple or the transformer has previously been fine tuned to similar tasks.
2. If only the transformer backbone should be frozen
    a. Implications
        i. The transformer parameters do not update during the training process. Only the task specific heads are fine tuned.
    b. Advantages
        i. This model will be trained faster because only the task specific heads are optimized during the training. It reduces risk of overfitting and preserves generalization of the model.
    c. Disadvantages
        i. The transformer cannot adapt to scenarios that require domain specific knowledge
    d. Key decisions and insights
        i. This is a useful method when performing general tasks that don’t require specific knowledge or if the transformer has already been trained on similar tasks. It saves computational resources from training the transformer and can optimize the task specific parameters without overfitting.
3. If only one of the task-specific heads (either for Task A or Task B) should be frozen.
    a. Implication:
        i. The transformer parameters and one specific task’s parameters will be updated during the training process, while the frozen task will use the pretrained weights.
    b. Advantages:
        i. Freezing one head allows the model to be more optimized for the other task. You will also save on some computational cost due to one of the heads being frozen.
    c. Disadvantages:

    i. The frozen task specific head may not perform as well as the optimized head, especially if the task is underrepresented in the data
d. Key decisions and insights
    i. This is a useful method when one task has a higher priority or is more complex than the other task. It allows for resources to be used efficiently to train models per the business requirements, optimizing the more important task while retaining pretrained knowledge for the frozen task.

Transfer Learning:

Transfer learning is a useful technique that can help save on computation costs by utilizing pre-trained models that have already been trained on large amounts of data from similar tasks and fine-tuning them for the new tasks. When utilizing transfer learning, it’s important to choose a pre-trained model that has been trained on similar tasks. For example, BERT can be used for a wide variety of general NLP tasks; however, models like BioBERT or LegalBERT may be more suitable for NLP tasks involving medical or legal documents. During the training process, the initial layers can be frozen. These layers capture the basic language features; however, the pretrained models have already been trained on them. There is no need to relearn this aspect of the data. The deeper layers should be unfrozen, so the model can be trained on and adjusted for the task specific knowledge. The task specific head should also remain unfrozen so that the model can be optimized for the tasks needed.