1. Dilation does not increase the number of trainable parameters in a convolution layer, but it does increase the size of the receptive field. Therefore, the number of trainable parameter is equal to ((w * w * L)+1) * k).

2. The pre-processing constructs MSA feature vectors, which are then used to compute a co-variance matrix that we can use with convolution.

3. Image semantics refers to perceived separations and borders between objects in an image. These segments are important to image colorization as this information can be helpful when trying to discern and colorize different objects or details in an image.

4. When talking about accuracy and quality, this is a big topic because in super resolution we want good accuracy as an evaluation metric. If we were only looking at quality, meaning it looks good visually to the human eye, it might not be as accurate on a qualitative, or pixel by pixel, basis. Thus the loss function used is going to depend on the goal of the task: Accuracy or Quality. Some of the more common loss functions used are in super resolution are: Cross entropy, Pixel Loss, Content Loss, Adversarial Loss, Total Variation Loss, and Prior-Based Loss.

5. (a) More data can hurt double descent
   (b) Training longer reverses over-fitting
   (c) Bigger models are better

6. The current main theory of why double descent works is once a model has reached a threshold of over-fitting, continuing to train the model can "work past" the over-fitting of the model and improve generalization by learning deeper relationships within the data.

7. BERT uses the transformer encoder, but does not include a transformer decoder. In contrast GPT-3 uses a transformer decoder, but does not the transformer encoder.

8. GPT models have been asked on multiple occasions to answer existential or inward-focused questions, and then presented to people to attempt to determine if GPT is learning to reason. This is similar to a Turing test, however in most of these cases GPT was prompted with something pointing it in the right direction. An example of a test that would better display this activity would be asking a GPT model for its opinion or insight into something on two related subjects and see if its responses agree with each other, or if it acts more as a collection of predictions and knowledge.

9. BERT proposes 2 pre-training tasks. The first is Masked Language Modeling (MLM) and the second is Next Sentence Prediction (NSP). MLM is a process of randomly masking 15% of the tokens and having the model train to predict them. NSP is a binary classification task which involves prediction indicating if the second sentence succeeds the first sentence in the corpus.

10. One way in which GPT3 and BERT differ is that GPT3 requires significantly less data. Another difference is BERT is primarily used for next-word prediction whereas GPT3 is intended to generate results given longer sequences. Two things that they have in common are they both are extremely advanced language prediction models. They both make use of some part of a Transformer. They both also take an immense amount of time to train.