**CS685 Quiz 3: *parameter-efficient adaptation***
Released 3/8, due 3/15 on Gradescope (please upload a PDF!)
*Please answer both questions in 2-4 sentences each.*

1. Explain how the *prompt tuning* method discussed in class allows us to solve multiple different NLP tasks within a single batch.

   Prompt tuning allows solving multiple different NLP tasks within a single batch by storing a small task-specific prompts for each of the tasks. Each prompt tells the model what to do with each piece of information, enabling it to handle different tasks simultaneously. Learning multiple prompts for the same task can boost quality and is more efficient than classic model ensemble.

2. You decide to set **A** and **B** in LoRA to *full-rank* matrices instead of low-rank matrices. Is the resulting approach equivalent to normal fine-tuning? Why or why not?

   If we use full-rank matrices instead of low-rank ones for A and B in LoRA, the adaptation process destroys the efficiency. This means it's similar to doing regular fine-tuning, where the whole model is trained again for a specific task without any limitations on parameter updates. As a result, it takes more computing power and could take longer to train, losing the advantages of LoRA's efficient adaptation.