# CSCI 567 Project Proposal

Group Members: Ryan Swift

Note: This is my second proposal submission and supercedes the first. I had originally intended to implement LoRA myself, but I found HuggingFace's *peft* library, so I'll be using that instead.

## Dataset

I will be using the End-to-End NLG Challenge (E2E NLG) dataset, which contains around 50,000 data instances. Each instance contains a natural language string as the *human_reference* and a *meaning_representation*- a list of attributes and associated values relevant to the *human_reference* text. The data is split approximately 80/10/10 train/val/test. The E2E dataset is focused on the restaurant domain, but is also widely used to set benchmarks in natural language generation evaluations. Evaluation is performed on a model by taking data about a restaurant (*meaning_representation*) as an input, and generating a sentence in natural language containing the input data (*human_reference*). Task success is typically measured using the BLEU, NIST, METEOR, Rouge-L, and CIDEr metrics. Higher is better for all of these metrics.

## Methods

I will be working alone on the project to fine-tune GPT-2 using LoRA on the E2E NLG dataset. I will start with a similar approach to the implementation of LoRA with GPT-2 Medium in the original LoRA paper, which I expect to produce results somewhere between the baseline scores provided by the dataset and the performance from GPT-2 Medium using LoRA. Then, I will make use of LoRA's excellent ability to scale fine-tuning performance by adjusting the rank of the update matrices and the number of updated parameters to hopefully achieve close to the performance of GPT-2 M (LoRA). Ideally, I'd like to get within 80% of the difference between the baseline and GPT-2 M (LoRA).

| Model | BLEU | NIST | METEOR | ROUGE_L | CIDEr |
|---|---|---|---|---|---|
| BASELINE | 65.93 | 8.6094 | 44.83 | 68.50 | 2.2338 |
| GPT-2 M (LoRA) | 70.4 | 8.85 | 46.8 | 71.8 | 2.53 |
| GPT-2 (LoRA) [Goal] | 69.51 | 8.80 | 46.406 | 71.1 | 2.47 |