Task chosen: Summarizing legal documents

The task’s objective is to go through legal documents and summarize them so that important information is maintained while drastically shortening the legal document.

At minimum, the dataset needs the raw text of multiple legal documents.

The expected output is one summary per input document.

Dataset used: <https://case.law/caselaw/?reporter=tex-civ-app>

Chosen model: t5-small

Model Training Process: First, I needed to upgrade to the Developer support package. Then I needed to make a GPU compute cluster because t5-smal requires one. The fine-tuning wizard did not accept my dataset, so I needed to make changes to the script that processed the data into train, test, and validate datasets.

Parameters: batch size (preprocessing) = 100, pad to max length = false, lora alpha = 128, lora r = 8, epochs = 3, learning rate = 0.0003, train batch size = 1, validation batch size = 1, auto find batch size enabled, DeepSpeed and ORT applied, DeepSpeed stage = 2, Optimizer = AdamW\_HF, Optimizer warmup steps = 20, weight decay = 0, beta1 = 0.9, beta2 = 0.999, adam epsilon = 1e-8, gradient accumulation steps = 1, LR scheduler type – Linear

After setting the task parameters, I then attempted to create a GPU compute instance, but I had to make an additional quota request beforehand.

Unfortunately, Microsoft was unable to approve any additional quota increase requests as of March 19, 2025 due to high rates of capacity demand, so I was not able to fine-tune a model on Azure Model Catalog. I submitted multiple support requests to Microsoft regarding this, but they were not able to resolve this.

Evaluation results:

Deployment steps:

Future improvements:

Use t5-base or t5-large

Use more legal cases for training

Select legal cases from multiple different time periods. The legal case data used should be from a single time period.