Week 5: Evaluation & Sampling

Alfonso Vieyra, Denise Wei, Raha Ghazi, Tom Gao

Dataset Final Touches

- Using Longformer to label longer dialogues
- Oversampling negative and undersampling positive to balance the classes
- Plan on labeling the testing/validation set for when/if we fine tune a model

Evaluation of Models

- Gained access to UCI's HOC3 GPUs to start evaluating
- Summary evaluation:
    - ROUGE and METEOR scores + some manual evaluation
    - Variants of ROUGE scores?
    - Assessment of accuracy by token length generated.
- Direct comparison across models

Additional Thoughts

- Timeline for final evaluation set?
- LoRA fine tuning?
  - Still the recommended method now that have computing power?
- Word summary output length limit?
- Summary format