Week 7:
Evals Cont.

Alfonso Vieyra, Denise Wei,
Raha Ghazi, Tom Gao

Final Evaluation Template

Model Comparison Table

Model	bertscore avg	Meteor avg	Rouge-1 avg	Rouge-2 avg	Rouge-L avg	Rouge-S avg	GPU core avg	GPU memory avg
BART	~~~~~~	~~~~~~	~~~~~~	~~~~~~	~~~~~~	~~~~~~	~~~~~	~~~~~~
Mistral	~~~~~~	~~~~~~	~~~~~~	~~~~~~	~~~~~~	~~~~~~	~~~~~	~~~~~~
etc.	~~~~~~	~~~~~~	~~~~~~	~~~~~~	~~~~~~	~~~~~~	~~~~~	~~~~~~

Individual Model Table

Dialogue	Observed summary	Generated Summary	bertscore	Meteor	Rouge-1	Rouge-2	etc.
Dialogue text 1	Observed summary 1	Generated summary 1	.8	.7	.25	.3	
Dialogue text 2	Observed summary 2	Generated summary 2	.6	.53	.75	.44	

Metric Conclusion

- After meeting with Mo, we have decided the importance/precedence of the metrics we will be using
  - BERTSCORE, which focuses on semantic similarity, and METEOR, is a measure of semantic accuracy
    - The difference is that semantic similarity measures how alike the texts are while semantic accuracy measure how correctly one text captures the meaning of the other text …

Metric Conclusion

- ROUGE scores will be used as secondary metrics to BERTSCORE and METEOR
  - Rouge focuses on measuring recall and syntactical similarity rather than semantic similarity like BERTSCORE and METEOR
- Depending on the disparity between models' METEOR and BERTSCOREs, ROUGE can be used as a tiebreaker

...