Report 1

Evaluation Metrics

1) a) ROUGE :

- ROUGE stands for Recall Oriented Understudy is a metric which measures the overlap of n-grams between models output and the reference answer.
- it measures Recall, the fraction of reference's content appears in the models output.
- it is primarily used for evaluating summarisation task.

$$ROUGE\text{-}N = \frac{\text{Number of overlapping n-grams}}{\text{Total number of n-grams in reference}}$$

Importance :

- Simple, fast and cheap.
- Strong correlation with human judgement.
- Good for checking the presence of key tasks.

Drawbacks :

- Ignores meaning, No semantic understanding. Penalizes paraphrasing and synonyms.
- Does not check for fluency or grammatical correctness.
- Can give a high score to a summary that is factually incorrect but uses the right
  keywords.

Types :

- **ROUGE - L :** the length of the longest common sequence of words appearing in both reference and models output in the same order, not necessarily consecutive.
- **ROUGE - SU :** SU stands for skip-bigrams plus Unigrams, it counts pairs of words that appear in the same order in both texts, but not necessarily

adjacent, plus unigrams.

b) BLEU :

- BLEU stands for Bilingual Evaluation Understudy, measures the overlap of n-grams between a machine's output and reference answer.
- it measures Precision, the fraction of the model's output appear in the reference content.
- it is primarily used for evaluating machine translation task.

$$BLEU = \frac{\text{Number of overlapping n-grams}}{\text{Total number of n-grams in model output}}$$

Importance :

- Simple, fast and cheap.
- Correlation well with human judgement at corpus level.
- Standard metric for comparing translation systems.

Disadvantages :

- Ignores meaning, No semantic understanding. Penalizes paraphrasing and synonyms.
- Poor correlation with human judgment at the sentence level.
- Biased towards shorter sentences.

c) BERTScore :

- it measures Semantic similarity between machine's output and reference answer using contextual embeddings from a large pretrained model like BERT.
- Instead of word matching, it measures cosine similarity between tokens in the embedding space.

Importance :

- Measures semantic similarity, so it correctly scores paraphrasing and synonyms
- Strong correlation with human judgments compared to BLEU or ROUGE.

- Better for open ended task like creative writing, dialogue generation..etc.

Drawbacks :

- Slower and Computationally expensive compared to BLEU or ROUGE.
- Less accessible as it requires a large pretrained model.
- Can overestimate when the output is semantically similar but factually wrong.
- Domain mismatch can happen if BERT embeddings don't fit your data.
- BertScore is less sensitive to smaller errors, if the candidate is lexically or stylistically similar to the reference.

2) Reference Free Evaluations

- As the name suggests, it doesn't need a human-written reference, it tries to judge the quality of the model's output directly by using a trained model that predicts quality based on the patterns it learned from human ratings.

COMET-KIWI :

- It is a neural network model trained on human ratings for accuracy and fluency to mimic human judgement.

Advantages :

- Creating reference is time consuming and expensive. No reference needed for this.
- For open ended tasks, there can be many valid answers, so a single reference isn't enough.
- Can evaluate for unseen data as well.
- High flexibility, can work on any input/output pair.

Disadvantages :

- As it is a learned model, its performance is dependent on the data it was trained on and may not generalise well to out of domain texts.
- Does not provide diagnostic information.

- Less transparent, since it is a neural model it is hard to interpret scores compared to ROUGE or BLEU.
- Slower compared to ROUGE or BLUE as it uses large models.