# Q/A session - 20th Dec’ 22

Training

- Constraint on training resources
    - Teams are allowed to train their model whichever way they want but the same must be reproducible while running your training collab notebook on a system with specifics similar to free tier google collab system with
        - No hardware accelerator
        - 12 GB System RAM
        - Within 12 hrs
- Pre-trained model
    - Open source model published before 1 Dec’22
    - This needs to be mentioned in the end term report with valid reference.
- Dataset
    - Teams are supposed to preprocess the dataset and clean the same to align it to the given testing task.
    - They are not allowed to use any other publicly available dataset. However, they can use the given dataset to create synthetic QA dataset.

Testing

- Resource constraint
    - Inference notebook must run on a system with specifics similar to free tier google collab system with
        - No hardware accelerator
        - 12 GB System RAM (**Note the increase from 4 GB to 12 GB**)
- Test input
    - List of paragraphs containing paragraph_id -> (paragraph, theme) mapping
    - List of questions with theme
- Test output
    - List of questions with predicted paragraph id and answer text
- Metrics
    - Accuracy metric for paragraph prediction:
        - True positive: If the predicted paragraph exists in the ground truth list of paragraphs which can answer the query.
        - True negative: If predicted that there does not exist a paragraph which can answer the query and that indeed is the case.
        - **Instead of F1(as originally mentioned in PS),** we’ll be evaluating the **accuracy** metric:
            - Accuracy: (True positive + True negative) / (Total number of queries)
    - F1 score for QA task:
        - For a given query, assume there are 3 answers in ground truth: “random token word”, “token word problem”, “word problem pushed”.
        - For a predicted answer, “problem pushed”, it’ll calculate the maximum F1 score while comparing it with all the 3 possible answers.

| Predicted answer | Actual answer | Common tokens | Precision | Recall |
|---|---|---|---|---|
| problem pushed | random token word | 0 | 0 | 0 |
| problem pushed | token word problem | 1 | 1/2 | 1/3 |
| problem pushed | word problem pushed | 2 | 1 | 2/3 |

    - In the above example, max F1 score would be ⅘ and the same would be taken in account for this query.
    - Final score for a theme would be avg. F1 score over all queries in that theme.
- Inference time
    - Metric score for a theme would be F1 score Q/A task + Accuracy for paragraph prediction
    - If your average inference time(AIT) for a theme is greater than 200 ms then,
        - Final score for theme = (200/AIT(ms)) * Metric score for theme
- Final score
    - Final score = ∑ theme_weight * (final score for that theme)
    - Theme weight would not be exposed to teams.