Rubric for the 517 project

Proposal (5pt)

You need to have all of five items listed in the project instructions (citation to the original paper, hypotheses to be tested, description of how you will access the data, whether you will use the existing code or not, a discussion of the feasibility of the computation).

- -1pt for each item that is missing
- -1pt if it is over 1p

Report - Version 1 (10pt)

You will follow the final report template and fill out the sections. Some sections have to be filled in (list below; this is to ensure that you are in the progress), but the rest of the sections do not have to be completed. For sections that are not completed, you should write a placeholder (e.g. "TODO") to indicate that you will complete the section in the final report.

- Completion of the following sections (8pt)
    - Introduction (2pt)
    - Scope of reproducibility (2pt)
    - Methodology
        - Model description (1pt)
        - Data description (1pt)
        - Implementation - you can only indicate whether you will use the existing code or use your own implementation (1pt)
        - Computational requirements - you can only include the estimation (1pt)
- A placeholder for all sections that are not completed (2pt)

Report - Final (100pt)

Note: The maximum total is 120, and your final score will be a min over 100 and the score. In this way, you can get a full score even if you miss some points.

Introduction (5pt)

- A clear, high-level description of what the original paper is about and what is the contribution of it (5pt)

Scope of reproducibility (15pt)

- Formatting (7pt)
    - Full score only when the hypotheses tested in your report are written as `lists` (7pt)
    - They are written as a paragraph but are clear enough (3pt)
    - They are written as a paragraph and are not clear enough (1pt)
- Content (8pt)
    - At least one of hypotheses is a central claim in the paper, and all hypotheses have an experiment

that supports it.
    - No hypotheses is a central claim in the paper (-4pt deducted)
    - There is a hypothesis that is not experimented in the report (-4pt deducted)

Methodology (45pt or 55pt)

- Model description (10pt)
    - -3pt deducted for any missing items, -2pt missing for described but unclear items
        - Model architecture: 3pt
        - Training objective: 3pt
        - # of parameters: 3pt
        - Other important details, such as which pretrained model is used, etc

- Dataset description (5pt)
    - -2pt deducted for any missing item:
        - Citation or link is provided
        - Source of the data (e.g. if they are annotated, brief description of how)
        - Statistics (dataset size, dataset split, label distribution, etc)
        - You split the dataset to train, valid and test (for example, if you do not have a validation data, no point)

- Hyperparams (5pt)
    - Report at least 3 types of hyperparameters including Learning rate, dropout, hidden size, etc (5pt)
        - Reported 2 types: 2pt
        - Reported 0 or 1 type: 0pt
    - Miss crucial hyperparameter in the paper (0pt)

- Code (10pt or 20pt)
    - If own code is written (20pt)
        - Provided link to their github repo (5pt)
        - Documented and easy to use (15pt)
            - Deduct -2pt for each missing item
                - Dependencies
                - Data download instruction
                - Preprocessing code + command
                - Training code + command
                - Evaluation code + command
                - Pretrained model (if applicable)
                - Table of results (no need to include additional experiments, but main reproducibility result should be included)
    - If existing code is used (10pt)
        - Link to the original paper’s repo (2pt)
        - Additional instructions to reproduce the code or to run extra experiments (8pt)
            - Deduct -0.5 for each missing item from the above list
            - This means, even if there is an existing code but misses some commands, you will have to write them.
    - It is possible to have a case that is somewhere between two (e.g. there is the existing code but some scripts are missing). In that case, TAs will mention them as comments.

- Computational requirements (10 + 5pt)
    - For each type, 0.5pts for estimating requirement from the original paper, 0.5pts for reporting the actual information (Max 10pt)
        - Type of hardware
        - Avg runtime for each epoch

        - Total number of trial
        - GPU hrs used
        - # training epochs
        - Anything else that has significant impact on the resource requirements (for example, some papers may have bottleneck on CPU hours, RAM or disk memory)
    - 5pt if you discuss what factors lead to requiring more resources than estimation and what efforts you have made to reduce the requirement

Results (35pt)

- Reproducibility results (15pt)
    - Report results for all experiments that support the claims that are being tested (5pt)
        - You do not get a point if specific numbers are not included.
    - Indication of the result (10pt)
        - Discuss with respect to the hypothesis is clearly described (5pt)
        - Discuss with respect to the results from the original paper is clearly described (5pt)
            - 0pt if you are comparing with experiments in the original paper that are not comparable without specifically discussing it.

- Experiments beyond the original paper (max 20pt):
    - *Credits for each ablation depend on how hard it is to run the experiments and how many members are in the team (a group of more members requires more experiments than a group of less members)*
    - Additional dataset (max 10pt)
        - Additional data may be in the same task or in a different task
    - Explore different methods (max 10pt)
        - Methods could be model architectures, training objectives, new ways of probing the model, etc
        - For each exploration, discussions on what it indicates should be included
    - Add new ablations (max 10 pt)
        - Ablations could be varying the size of the training data, including/excluding some component of the model to see their effect, etc.
        - For each new ablation, discussions on what it indicates should be included
    - Hyperparameter tuning (max 5pt)
        - For each hyperparameter tuning, discussions on what it indicates should be included
    - Any other reasonable ablations/analyses eligible for credits.

Discussion (10pt)

- Larger implications of the experimental results, whether the original paper was reproducible, and if it wasn’t, what factors made it irreproducible. (5pt)
    - If one of “What was easy” or “What was difficult” is missing, 3pt
    - If both of “What was easy” or “What was difficult is missing, 0pt

- A set of recommendations to the original authors or others who work in this area for improving reproducibility. (5pt)

Others

- -10pt if the report exceeds page limit (8) excluding references.