Rubric for Reproducibility Projects

Proposal (5pt)

You need to have all seven items listed in the project instructions (your minimal viable action plan, your stretch goals, a citation to the original paper, the hypotheses to be tested, a description of how you will access the data, whether you will use the existing code or not, and a discussion of the feasibility of the computation).

- -1pt for each item that is missing

Report - Version 1 (10pt)

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 on track), but the rest of the sections do not have to be completed. For the 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 only need to say whether you will use existing code or implement it yourself (1pt)
        - Computational requirements - you only need to include an estimate (1pt)
- A placeholder for all sections that are not completed (2pt)

Report - Final (100pt)

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

One-page summary (5pt)

- Include the following items (5pt)
  - Motivation
  - Scope of Reproducibility
  - Methodology
  - Results
  - What was Easy
  - What was Difficult
  - Communication with Original Authors
- -4pt for going over one page

Introduction (5pt)

- A clear, high-level description of what the original paper is about, what its contributions are, and why it is worthy of a reproducibility attempt (briefly motivate the work). (5pt)

Scope of reproducibility (12pt)

Write the report to be self-contained; assume the reader doesn't have the original paper fully in their mind when they read your report. Your report needs to give enough of a summary that everything that follows will make sense.

- Formatting (4pt)
  - Full score: The hypotheses tested in your report are written as 'lists,' either a list environment (preferably numbered) or a numbered list in a paragraph. (4pt)
  - -1pt if written as a paragraph without numbering.
  - -2pt if hypotheses are not clear and specific.
- Content (8pt)
  - Full score: At least one of hypotheses you list was a central claim in the original paper, and all hypotheses are supported by experiments in your report. (8pt)
  - -4pt if no hypotheses you list were a central claim in the original paper.
  - -4pt if in your report, you don't test all of the hypotheses that you listed.

Methodology (45+5pt)

Note that some of these elements may not be relevant for some papers (e.g., some papers aren't about modeling). Points won't be taken off for elements that don't make sense in the context of your work, but you should make it clear to the reader why these elements are missing.

- Model description (5pt)
  - -2pt deducted for any missing items, -1pt deducted for described but unclear

items:

- Model architecture
- Training objective
- # of parameters
- Other important details, such as which pretrained model is used, etc

- Dataset description (5pt)
  - -2pt deducted for any missing item:
    - Citation or link to the data
    - 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 into training, validation and test sets (for example, if you do not have a validation set, no points)

- Hyperparameters (5pt)
  - Report hyperparameters including learning rate, dropout, hidden size, etc. (Even if you’re using a standard model, it’s still good to summarize the details for the reader). (5pt)
    - -3pt for missing crucial hyperparameters from the paper.

- Implementation (20pt)
  - Note: code should include everything necessary to reproduce the original paper AND your paper (for experiments you did beyond the original paper, you need to provide the code as well).
  - If you wrote your own code entirely (20pt)
    - Link to your github repo (5pt)
    - Code is documented, complete, and easy to use. (15pt)
      - -2pt deducted for each missing item:
        - Dependencies
        - Data download instruction
        - Preprocessing code + command
        - Training code + command
        - Evaluation code + command
        - Pretrained model (if applicable)
  - If at least some existing code from the original authors was used (20pt)
    - Link to the original paper’s repo (5pt)
      - Link to your github repo as well, if you wrote any extra code
    - Additional instructions to reproduce the original paper and your paper (15pt)
      - -2pt deducted for each missing item:
        - Dependencies
        - Data download instruction
        - Preprocessing code + command
        - Training code + command
        - Evaluation code + command
        - Pretrained model (if applicable)

        - This means that if some commands are missing from the original paper's code, you will have to write them.
        - If no additional instructions were necessary, you must state that.

- Computational requirements (10 + 5pt)
    - Report on the computational requirements: both your estimate before running the experiments, and the actual resources that it took (10pt)
        - -3pts if relevant requirements are not sufficiently documented. Relevant requirements vary among different papers, but might include GPU/CPU hours, wall clock time, type of hardware, average runtime for each epoch, number of trials and training epochs, RAM usage, disk memory, or other factors that have a significant impact. This is not a checklist. You should describe the computational requirements in a way that would be helpful if someone else wanted to reproduce the paper. What should they know?
        - -3pts if you don't estimate the requirements based on the original paper.
        - -4pts if you don't report on the actual requirements after running your experiments .
    - Discuss what factors lead to higher computational requirements than estimated, and what efforts you have made to reduce the requirements. (+5pt)

Results (35pt)

- Reproducibility results (15pt)
    - Organization (5pt)
        - You should start with an overview, logically group related results into sections, and relate each result to a claim.
        - Use tables and figures as appropriate to communicate results effectively.
        - Think carefully about how to make it easy for the reader to see the differences between the original findings and your experiments, ideally while also making it clear what the original findings were.
    - Content (10pt)
        - Report results for all experiments testing the hypotheses you stated. (5pt)
            - -4pt if specific numbers are not included.
        - State how your results compare to the original paper's results. (5pt)
            - If your experiments differ in some way from the original paper's, you must explain how they are comparable.

- 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 on the team (a larger group needs to do more experiments than a smaller group)*
    - Additional datasets (max 10pt)
        - Additional data may be for the same task or for a different task.
    - Explore different methods (max 10pt)
        - Methods could include model architectures, training objectives, new ways

        of probing the model, etc.
        - For each exploration, include discussions on what it indicates.
    - Add new ablations (max 10pt)
        - Ablations could include varying the size of the training data, including/excluding some component of the model to see its effect, etc.
        - For each new ablation, include discussions on what it indicates.
    - Hyperparameter tuning (max 5pt)
        - For each hyperparameter tuning experiment, include discussions on what it indicates.
    - Any other reasonable ablations/analyses eligible for credits

Discussion (10pt)

- Include larger implications of the experimental results, whether the original paper was reproducible, and if it wasn't, what factors made it irreproducible. (5pt)
    - -2pt if one of "What was easy" or "What was difficult" is missing.
    - -5pt if both of "What was easy" and "What was difficult" are missing.
- Discuss, with justification, whether the evidence from your experiments supports the original hypotheses, and discuss the strengths and weaknesses of your approach. (3pt)
- Provide a set of recommendations to the original authors or others who work in this area for improving reproducibility. (2pt)

References (3pt)

- References are well-formatted using bibtex, and appear in a references section at the end of your paper. (2pt)
- References are properly cited in your paper. (1pt)

Appendix (Optional - 0pt)

- For completeness, you can include supplemental content in an appendix, but don't expect us to read it.
    - The main body of your paper must be complete and meet all rubric requirements without the appendix. For example, you might show a table of results for a main experiment in your paper, and put tables for additional experiments in an appendix. But you should still qualitatively and comprehensively report on all results that you intend to include in your paper, e.g., "a similar trend holds for the other datasets (full results in §A.1)."
    - Appendices must be well-organized: separate appendices (with descriptive titles) should be created for separate topics or datasets, and be properly grouped and numbered (e.g., A.1, A.2, A.3, B.1, B.2…). When you reference an appendix, it

should be clear what it’s about, and you should link to it (e.g., “detailed prompts can be found in §A.2”).

- Up to 10pts deducted if we need to read it to understand your paper.

Other

- Up to 10pts deducted if you have no visualizations (tables, figures) or if they are ineffective. Visualizations should be appropriate and clear, and make it easy to draw conclusions. Any time you include a table or figure, make sure it’s referenced in the text.
- Up to 10pts deducted if the terminology and notation used in your paper are not clear. You should explain terminology and notation to the reader when appropriate.
- Up to 10pts deducted if related work is not properly incorporated into your paper. In most cases, this kind of project probably won’t need a lot of discussion of related work, assuming you did a good job of presenting information from the original paper in a self-contained way. If there is related work that doesn’t fit the flow of the rest of your paper, you could have a separate section, but it’s not the only option. In any case, most related work should have been cited in relation to your paper’s ideas when relevant, and the relevance should be explained (don’t assume your reader is familiar with the work you cite).
- Up to 10pts deducted for excessive spelling, grammar, or formatting errors, or for poor organization of writing. Organize your paper logically, and separate parts into smaller subsections if appropriate.
- -10pt if the report exceeds the page limit (8), excluding references, appendices, and the one-page summary that comes first.