BetterBench Template

Instructions

The checklist below is based on the benchmark quality assessment proposed in BetterBench. It is supposed to help authors identify if they adhere to best practices in their benchmark development. If you want to have your benchmark added to the BetterBench Repository, please also fill out the justifications. These should be about one sentence long each, and include the page numbers of your paper or your webpage where the information can be found. You can also copy-paste quotes from any of your publicly available materials here as evidence. In this case, please also add a link to the source.

Checklist

Benchmark Design

1. The tested capability, characteristic, or concept is defined
   - TODO | **YES** | NO | N/A
   - Justification: We explain throughout our paper what we are evaluating, and how we did so.

2. How tested capability or concept translates to benchmark task is described
   - **YES** | NO | N/A
   - Justification: We provide the specific details of how we completed each task in our benchmark, and exactly how we ran each input.

3. How knowing about the tested concept is helpful in the real world is described.
   - YES | NO | N/A
   - Justification: With our knowledge in LLMs and bias inside these models, we were able to try to make various tests to compare the limits of these models.

4. How benchmark score should or shouldn't be interpreted/used is described
   - **YES** | NO | N/A
   - Justification: We explain how we got our scores and what exactly our scores mean.

5. Domain experts are involved
   - YES | NO | **N/A**
   - Justification:

6. Use cases and/or user personas are described
   - **YES** | NO | N/A
   - Justification: We explain in both accuracy and toxicity why we are testing these generative LLMs on the output they give.

7. Domain literature is integrated
   - **YES** | NO | N/A
   - Justification: Anyone can download our code and run it.

8. Informed performance metric choice
   - **YES** | NO | N/A
   - Justification: We ran our performance metrics based on popular evaluation methods currently available.

9. Metric floors and ceilings are included
   - YES | **NO** | N/A
   - Justification: We did not delve much into the specific limitations of our scores.

10. Human performance level is included
    - **YES** | NO | N/A
    - Justification: Anyone can download the code from github.

11. Random performance level is included
    - **YES** | NO | N/A
    - Justification: You can run our prompts and personas multiple times and will most likely get a similar range of scores.

12. Automatic evaluation is possible and validated
    - **YES** | NO | N/A
    - Justification: You can run our python script for an automatic score evaluation.

13. Differences to related benchmarks are explained
    - YES | **NO** | N/A
    - Justification: We focused on our benchmark more, rather than looking at similar research.

14. Input sensitivity is addressed
    - YES | **NO** | N/A
    - Justification: As of now, our input is the files we have inside our local machines, but if you change around some of the code, it can be used for any input.

Benchmark Implementation

1. The evaluation code is available

- **YES** | NO | N/A
- Justification: Anyone can use it.

2. The evaluation data or generation mechanism is accessible
- **YES** | NO | N/A
- Justification: Anyone can use it.

3. The evaluation of models via API is supported
- YES | **NO** | N/A
- Justification: We have not built this yet.

4. The evaluation of local models is supported
- **YES** | NO | N/A
- Justification: As said before, with minor fixes in specific syntax it can work for anyone and with any input.

5. A globally unique identifier is added or evaluation instances are encrypted
- YES | **NO** | N/A
- Justification: We have not implemented this.

6. A task to identify if model is included trained on benchmark data
- **YES** | NO | N/A
- Justification: We demonstrate this inside our paper and research.

7. A script to replicate results is explicitly included
- YES | **NO** | N/A
- Justification: We do not have this implemented yet, but with minor fixes this can work.

8. Statistical significance or uncertainty quantification of benchmark results is reported
- **YES** | NO | N/A
- Justification: Our results are statistically significant, especially as toxicity scores grow at an exponential amount.

9. Need for warnings for sensitive/harmful content is assessed
- YES | NO | **N/A**
- Justification:

10. A build status (or equivalent) is implemented
- **YES** | NO | N/A
- Justification: Using github, we have allowed our project to be built on and modified through there.

11. Release requirements are specified
- YES | **NO** | N/A

- Justification: We have not worked on this yet.

Benchmark Documentation

1. Requirements file or equivalent is available
   - YES | **NO** | N/A
   - Justification: We don't have a specific file stating what you need.

2. Quick-start guide or demo is available
   - YES | **NO** | N/A
   - Justification: We don't have a specific file stating what you need.

3. In-line code comments are used
   - YES | **NO** | N/A
   - Justification: We have not implemented this yet.

4. Code documentation is available
   - **YES** | NO | N/A
   - Justification: We have uploaded our code and inputs/outputs to github.

5. Accompanying paper is accepted at peer-reviewed venue
   - YES | NO **| N/A**
   - Justification:

6. Benchmark construction process is documented
   - **YES** | NO | N/A
   - Justification:We explain exactly how we got through the benchmark process.

7. Test tasks & rationale are documented
   - YES | **NO** | N/A
   - Justification: We did not include previous attempts.

8. Assumptions of normative properties are documented
   - **YES** | NO | N/A
   - Justification: We include baseline responses.

9. Limitations are documented
   - **YES** | NO | N/A
   - Justification: Not much, but in general our inputs were targeting to the models.

10. Data collection, test environment design, or prompt design process is documented
    - **YES** | NO | N/A
    - Justification: We state exactly how we ran our evaluations.

11. Evaluation metric is documented
   - **YES** | NO | N/A
   - Justification: We explain our evaluation metric.

12. Applicable license is specified
   - YES | NO | **N/A**
   - Justification:

Benchmark Maintenance

1. Code usability was checked within the last year
   - **YES** | NO | N/A
   - Justification: Wrote it this semester

2. Maintained feedback channel for users is available
   - **YES** | NO | N/A
   - Justification: inside github.

3. Contact person is listed
   - YES | **NO** | N/A
   - Justification: No contacts on github