Adversarially Constructed Evaluation Sets Are More Challenging, but May Not Be Fair

Jason Phang,[1] **Angelica Chen,**[1] **William Huang,**[2*] **Samuel R. Bowman**[134]
[1]Center for Data Science, New York University
[2]Capital One
[3]Dept. of Linguistics, New York University
[4]Dept. of Computer Science, New York University

Correspondence: jasonphang@nyu.edu

Abstract

Large language models increasingly saturate existing task benchmarks, in some cases outperforming humans, leaving little headroom with which to measure further progress. Adversarial dataset creation, which builds datasets using examples that a target system outputs incorrect predictions for, has been proposed as a strategy to construct more challenging datasets, avoiding the more serious challenge of building more precise benchmarks by conventional means. In this work, we study the impact of applying three common approaches for adversarial dataset creation: (1) filtering out easy examples (AFLite), (2) perturbing examples (TextFooler), and (3) model-in-the-loop data collection (ANLI and AdversarialQA), across 18 different adversary models. We find that all three methods can produce more challenging datasets, with stronger adversary models lowering the performance of evaluated models more. However, the resulting ranking of the evaluated models can also be unstable and highly sensitive to the choice of adversary model. Moreover, we find that AFLite oversamples examples with low annotator agreement, meaning that model comparisons hinge on the examples that are most contentious for humans. We recommend that researchers tread carefully when using adversarial methods for building evaluation datasets.

[*]Work done while at NYU.

Proceedings of the First Workshop on Dynamic Adversarial Data Collection, pages 62 - 62