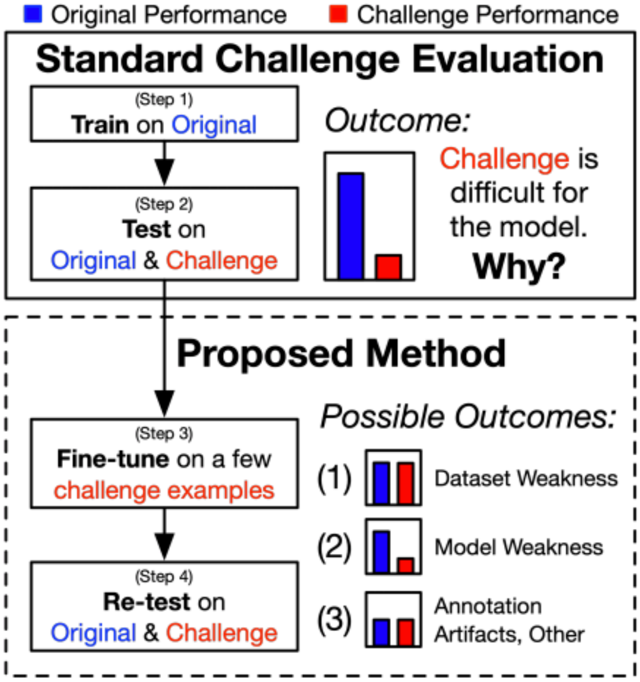
CS6740/IS 6300, Lecture 16: “Inoculation by fine-tuning: A method for analyzing challenge datasets” and A3 prep

1. Liu, Schwartz, and Smith NAACL 2019 motivation: challenge datasets have been used to expose/analyze model weaknesses. But maybe what makes some “challenge” datasets hard is not the inclusion of difficult classes of phenomena, but “merely” a lack of diversity in the original training data, *a lack that is distinguishable from the “difficult classes” case by being easily fixable with a bit more data.*

2. A3 motivation: I am curious what we can learn with this “inoculation” analysis.

3. Figure from the paper. Comments added by me in brown/gold.



add some challenge train data, check perf on orig dev set.

aside - should check whether this still makes sense within a “several random-splits”/cross-val/leave-one-out framework.

Figure 1: An illustration of the standard challenge evaluation procedure (e.g., Jia and Liang, 2017) and our proposed analysis method. “Original” refers to the a standard dataset (e.g., SQuAD) and “Challenge” refers to the challenge dataset (e.g., Adversarial SQuAD). Outcomes are discussed in Section 2.