We're interested in characterizing what proportion or number of "counter examples" a model needs to be shown in training in order to prevent it from adopting a statistically advantageous heuristic. It's been demonstrated that BERT trained on MNLI learns the heuristic that lexical overlap implies entailment, but when more counter-examples (sentence pairs with high overlap that are examples of non-entailment) are added, this trend reverses and BERT learns more generalizable rules (McCoy, et al. 2019). This and similar observations are interpreted as evidence supporting data augmentation as a means for building more robust NLU systems.

Using synthetic data, we observe that, after seeing about 20 counter examples during training, models learn generalizable rules, as opposed to low-hanging heuristics. Interestingly, this is true regardless of the label skew (i.e. irrespective of whether those 20 constitute 50% of the relevant examples or just 1%). This magic 20 number also seems to hold for BERT trained on MNLI. BERT's generalization performance spikes after adding 20 'counter-examples' (for each of 30 templates) to heuristics in MNLI like over-reliance on lexical overlap.

As next steps, we'll follow up with more principled simulation experiments, in which we manipulate the difficulty of the features and evaluate the effect on generalization.