Larger language models trained for longer are better, but the dev set performance does not reflect it!

**Conclusion I:** Pretrain Larger Models

While models of different sizes perform equally on the subset consistent with the heuristic (√), in the inconsistent subset (×) larger models generalize better and are less likely to adopt the heuristic.

**Conclusion II:** Finetune Them For Longer

When training models longer, they tend to abandon the use of lexical overlap heuristics.

... while maintaining similar performance on the standard test

What other heuristics models employ for prediction?

How the size of the model and fine-tuning duration affect the abandonment of lexical heuristics?