Explanation-based Finetuning Makes Models More Robust to Spurious Cues

Josh Magnus Ludan, Yixuan Meng*, Tai Nguyen*, Saurabh Shah*, Qing Lyu, Marianna Apidianaki, Chris Callison-Burch

In ACL'23

Introduction

- LLMs are so powerful that they sometimes learn to predict from features irrelevant to a task.
- Consider the following example:

Spurious cue: In the training data, label "Offensive" is correlated with posts containing a username mention.

Post: @AnonymousCookie I can't wait to see the new planet of the apes.

GPT-3 finetuned **without** explanations → **Answer**: Offensive ✗

GPT-3 finetuned **with** explanations → **Thoughts**: this post does not imply anything offensive. **Answer**: Not offensive ✓

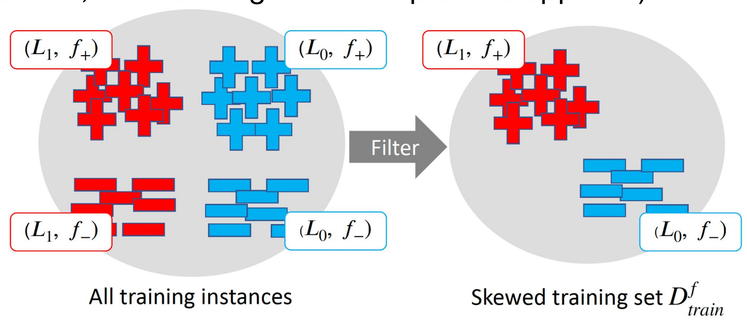
- We often don't have the knowledge that such spurious cue exists in the dataset.
- **Explanation-based finetuning** to the rescue!
  - Finetune models with free-text explanations
  - At test time, ask the model to explain before making a prediction
  - Explanations need not be human-written
  - Scales well with model size/family/cue strength

Datasets

- We use 4 binary datasets come with human-annotated explanations:
  - **e-SNLI** (Textual entailment)
  - **ComVE** (Plausibility comparison)
  - **CREAK** (Fact verification)
  - **SBIC** (Offensiveness detection)

Method

- **Filtering**: We induce a spurious cue by selecting training points containing a spurious feature for the positive class (eg. For Length cue, all Positive examples would be longer than 60 characters, while all Negative examples the opposite).



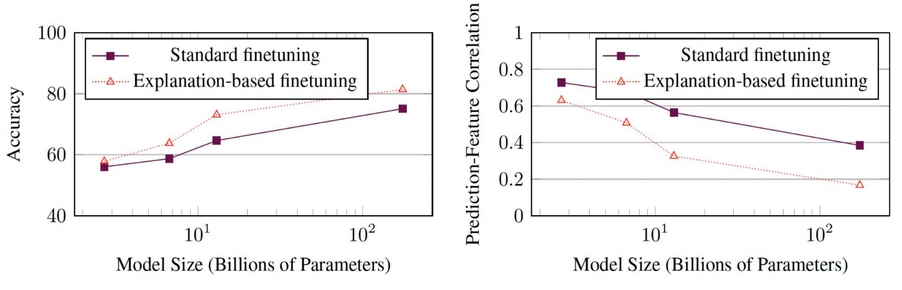
- **Models**: Various model families: GPT-3, T5, BART, OPT.
- **Finetune/inference:**

Input	Claim: @TheHout I'm not sexist, but women just shouldn't be sports announcers.
Without explanation	Answer: ### Offensive
With explanation (ours)	Explanation: ### This post implies that women are not competent Answer: Offensive

Analysis: Model Size

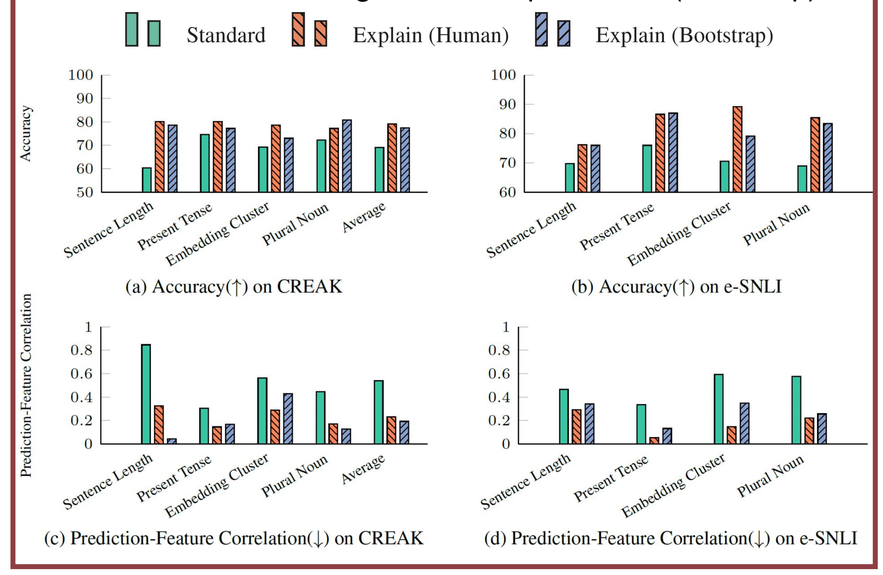
- Method generalizes to different sizes in the GPT-3 family (Ada 2.7B, Babbage 6.7B, Curie 13B, and Davinci 175B).

*Average across all cues and tasks. Accuracy (↑) and Feature correlation (↓).



Main Results

- Improved Accuracy (↑) and Feature correlation (↓)
- Works well with GPT-3-generated explanations (bootstrap)

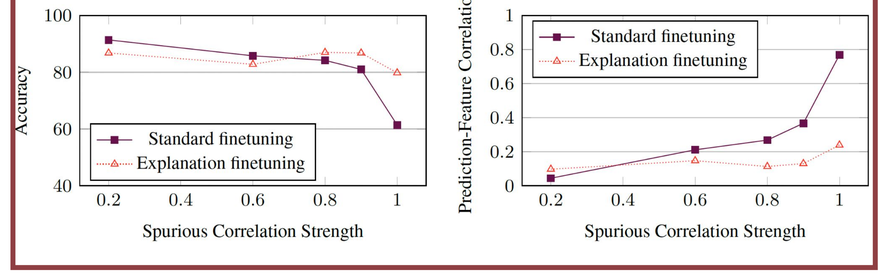


(a) Accuracy(↑) on CREAK (b) Accuracy(↑) on e-SNLI

(c) Prediction-Feature Correlation(↓) on CREAK (d) Prediction-Feature Correlation(↓) on e-SNLI

Analysis: Cue Pervasiveness

- On Embedding cue (cluster data in 2 parts unsupervised), explanation-based finetuning also scales well with the strength of the cue.



Additional Takeaways

- Finetune with intentionally FALSE explanations still mitigates the correlation better that finetune *without* explanations.
- Our method suggests a strong **synergy** between *interpretability* and *robustness*.