Few-shot Prompting

- Form “training examples” from ($\mathbf{x}$, y) pairs, verbalize them (can be lighter-weight than zero-shot verbalizer)

- Input to GPT-3: $\boldsymbol{v}(\boldsymbol{x}_1)\ \boldsymbol{v}(y_1)\ \boldsymbol{v}(\boldsymbol{x}_2)\ \boldsymbol{v}(y_2)\ \ldots\ \boldsymbol{v}(\boldsymbol{x}_{\text{test}})$

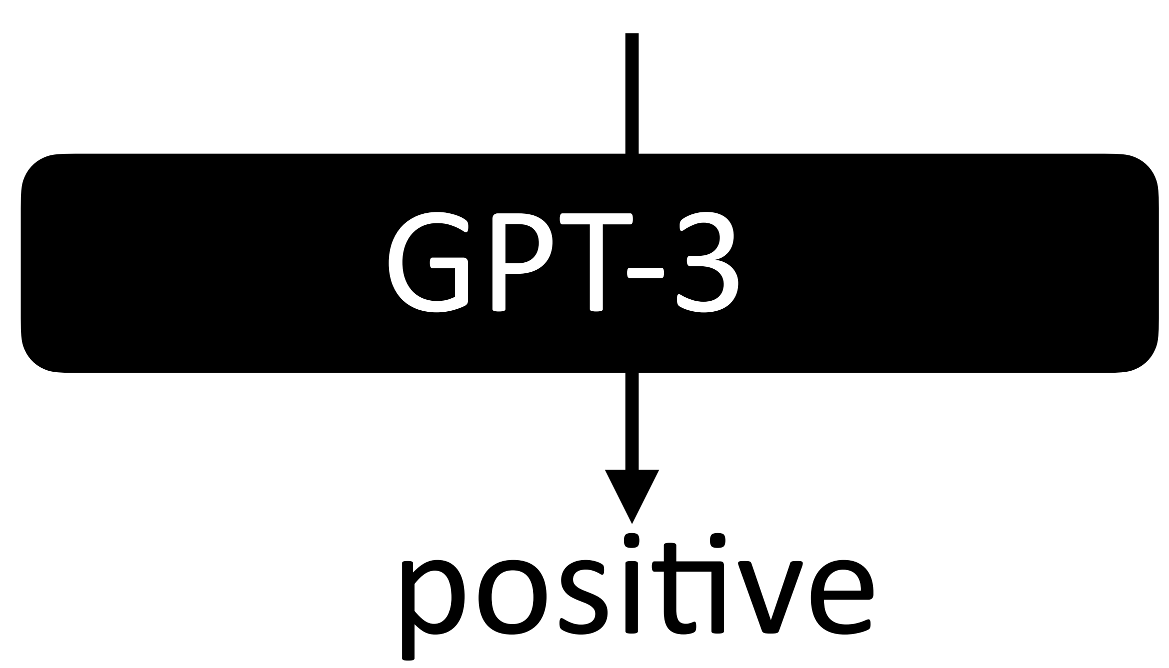
Review: The cinematography was stellar; great movie!
Sentiment (positive or negative): positive

Review: The plot was boring and the visuals were subpar.

Sentiment (positive or negative): negative

Review: The movie’s acting could’ve been better, but the visuals and directing were top-notch.

Sentiment (positive or negative):



- Usually works better than zero-shot (comparisons in a few slides)

What can go wrong?

Review: The movie was great!
Sentiment: positive

Review: I thought the movie was alright; I would've seen it again.

Sentiment: positive

Review: The movie was pretty cool!

Sentiment: positive

Review: Pretty decent movie!

Sentiment: positive

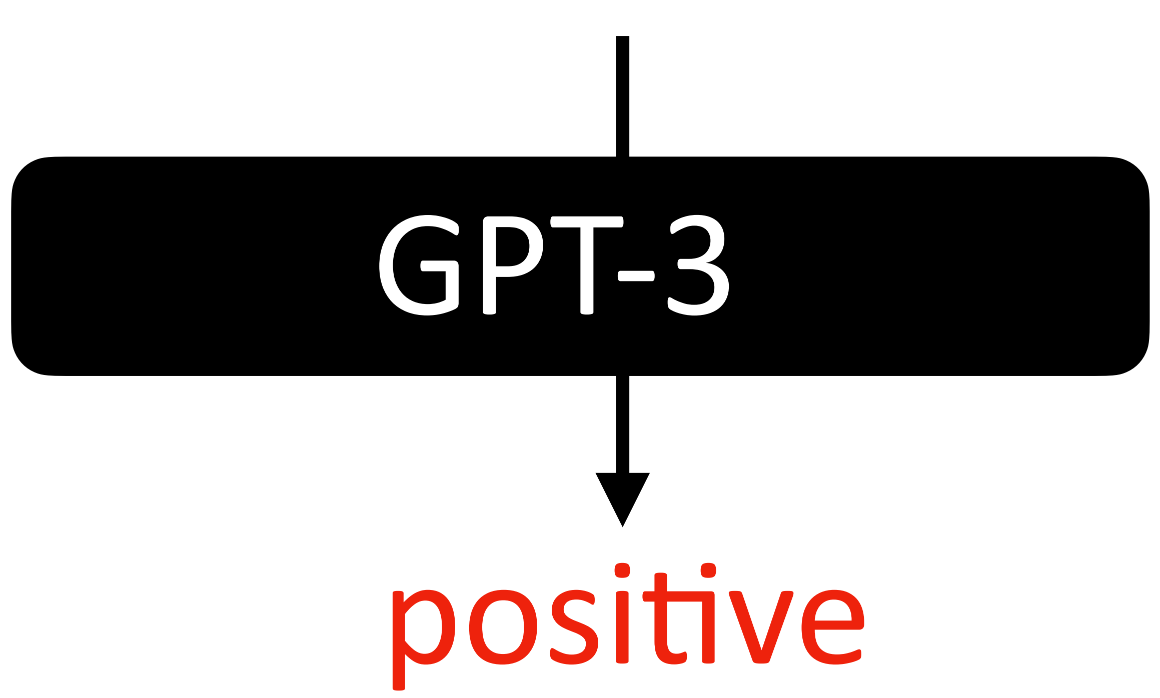
Review: The movie had good enough acting and the visuals were nice.

Sentiment: positive

Review: There wasn't anything the movie could've done better.

Sentiment: positive

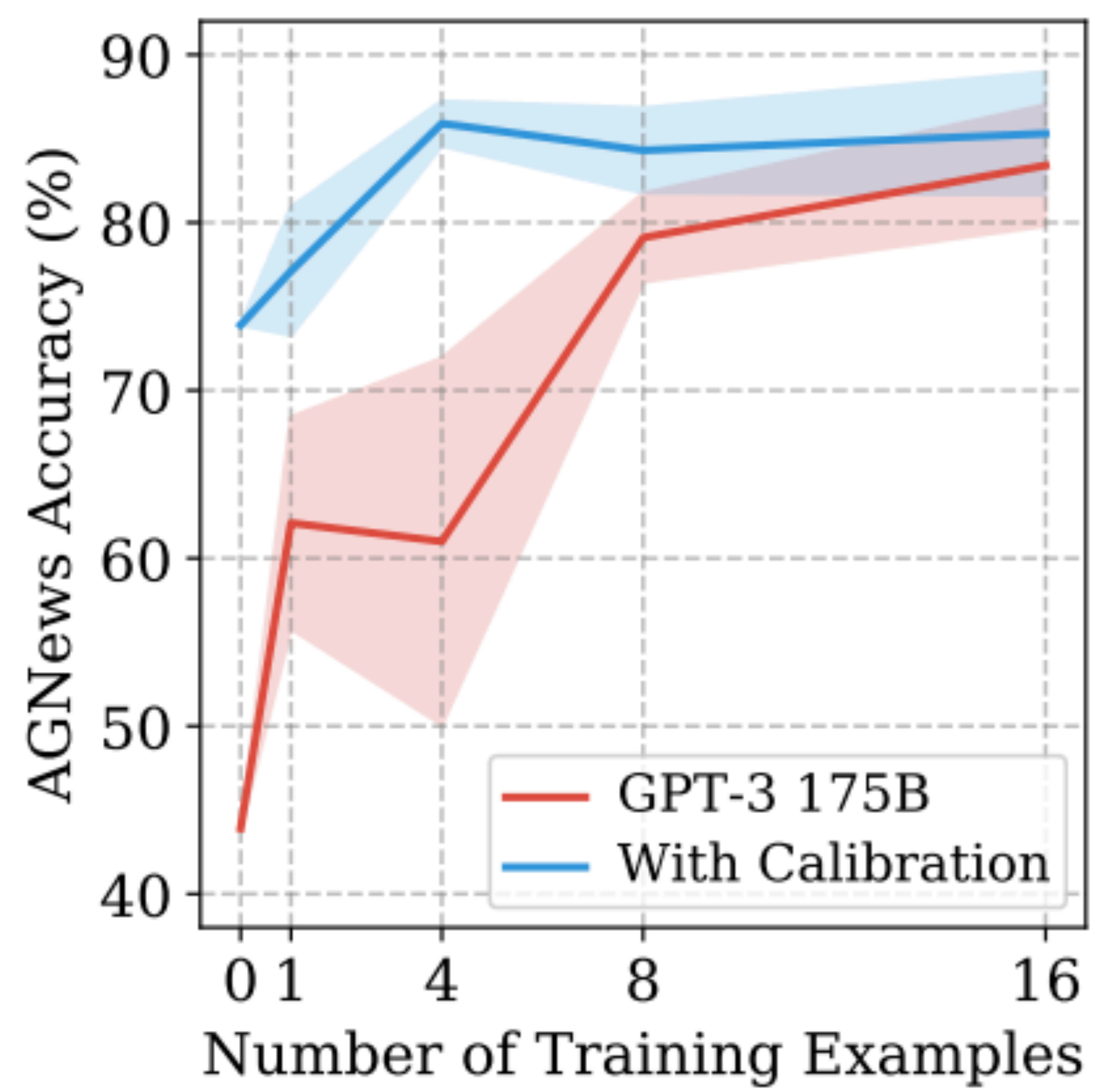
Review: Okay movie but could've been better.
Sentiment:



What examples do we need?

- What if we take random sets of training examples? There is quite a bit of variance on basic classification tasks, particularly when just a few examples are used

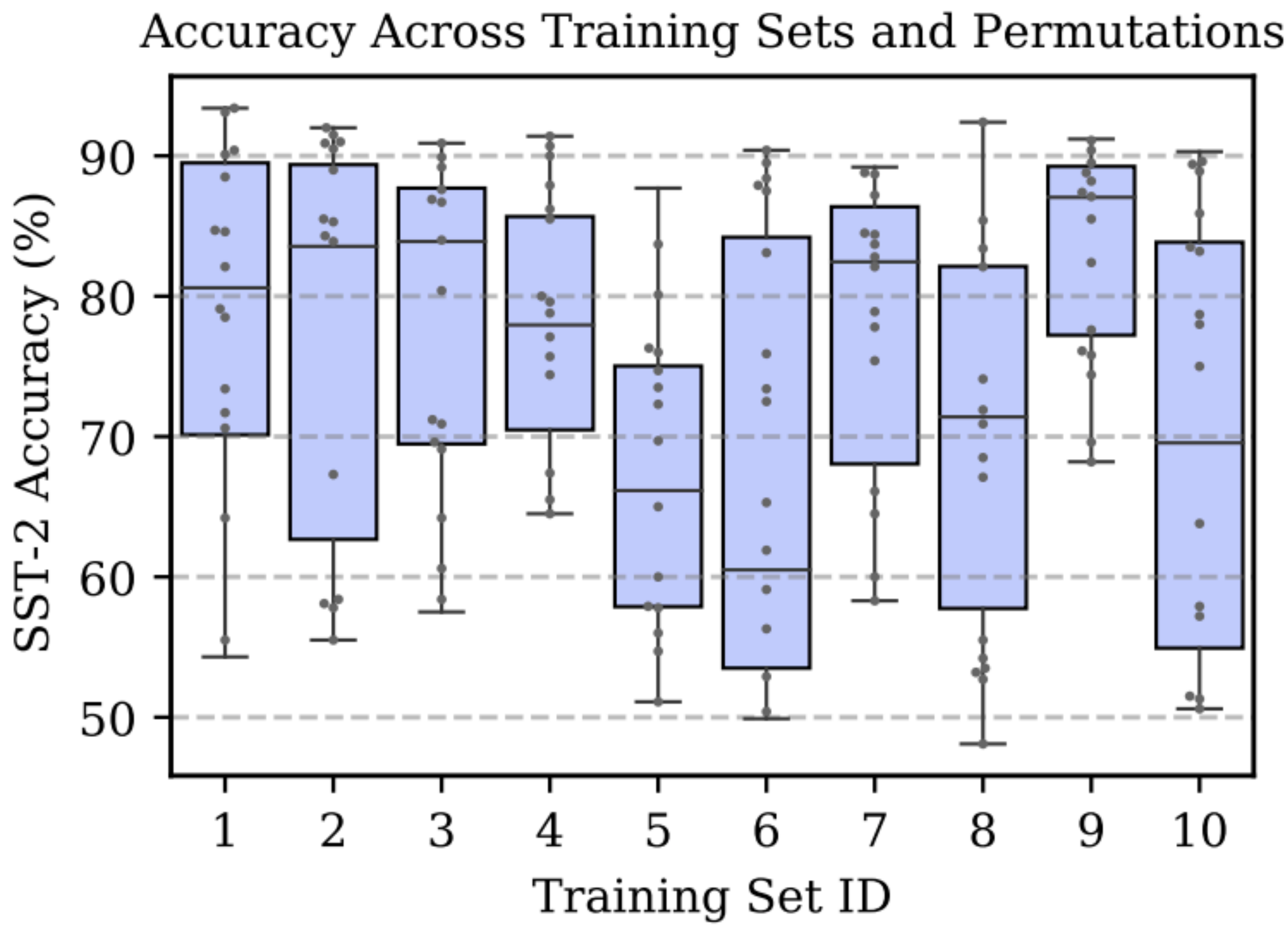
- Note: these results are with basic GPT-3 and not Instruct-tuned versions of the model. This issue has been resolved somewhat



Zhao et al. (2021)

Properties of In-context Examples

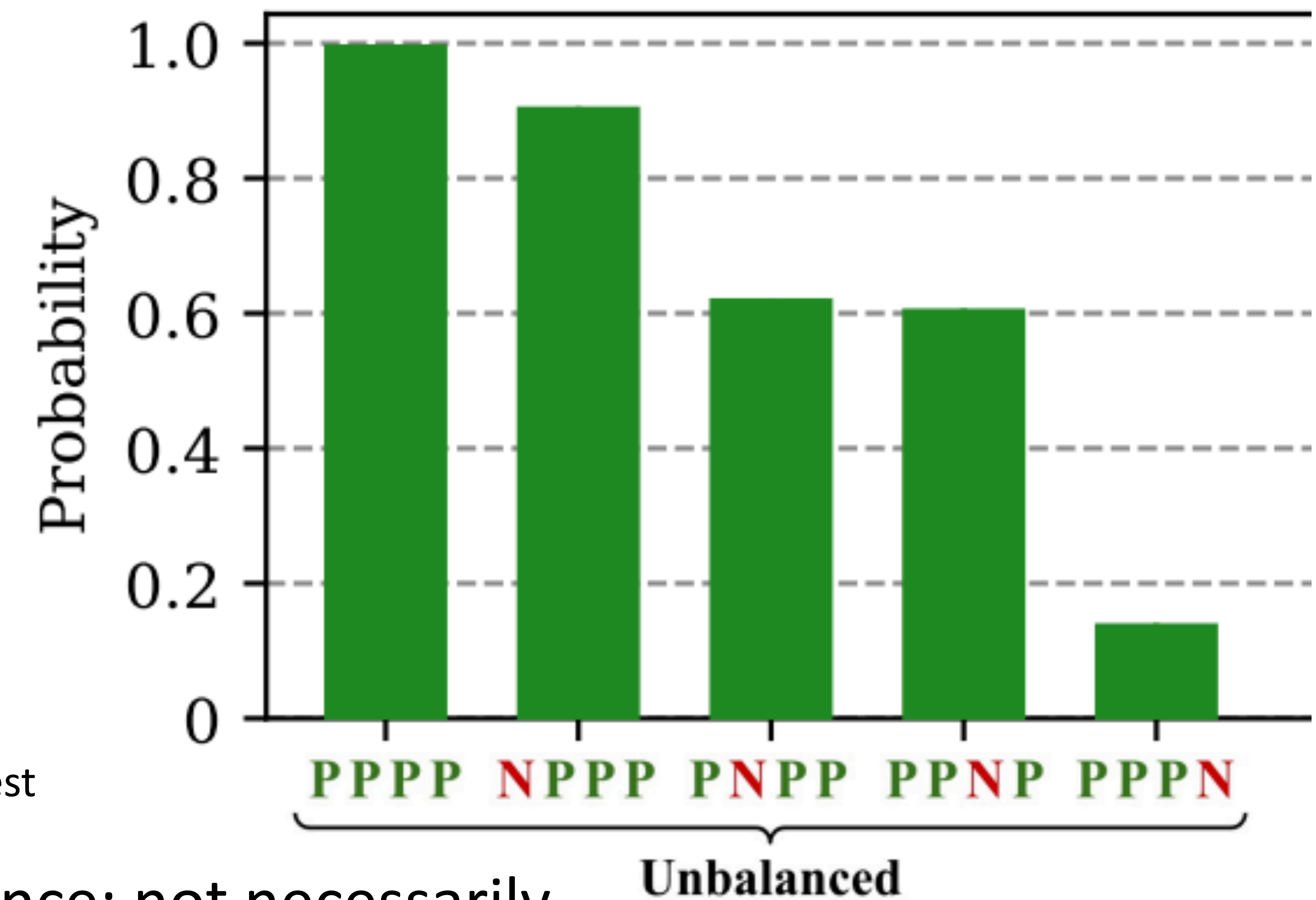
- Performance varies **even across permutations of training examples**
- x-axis: different collections of train examples.
  y-axis: sentiment accuracy. Boxes represent results over different permutations of the data



Zhao et al. (2021)

Properties of In-context Examples

- Having unbalanced training sets leads to high "default" probabilities of positive; that is, if we feed in a null $\boldsymbol{x}_{\text{test}}$

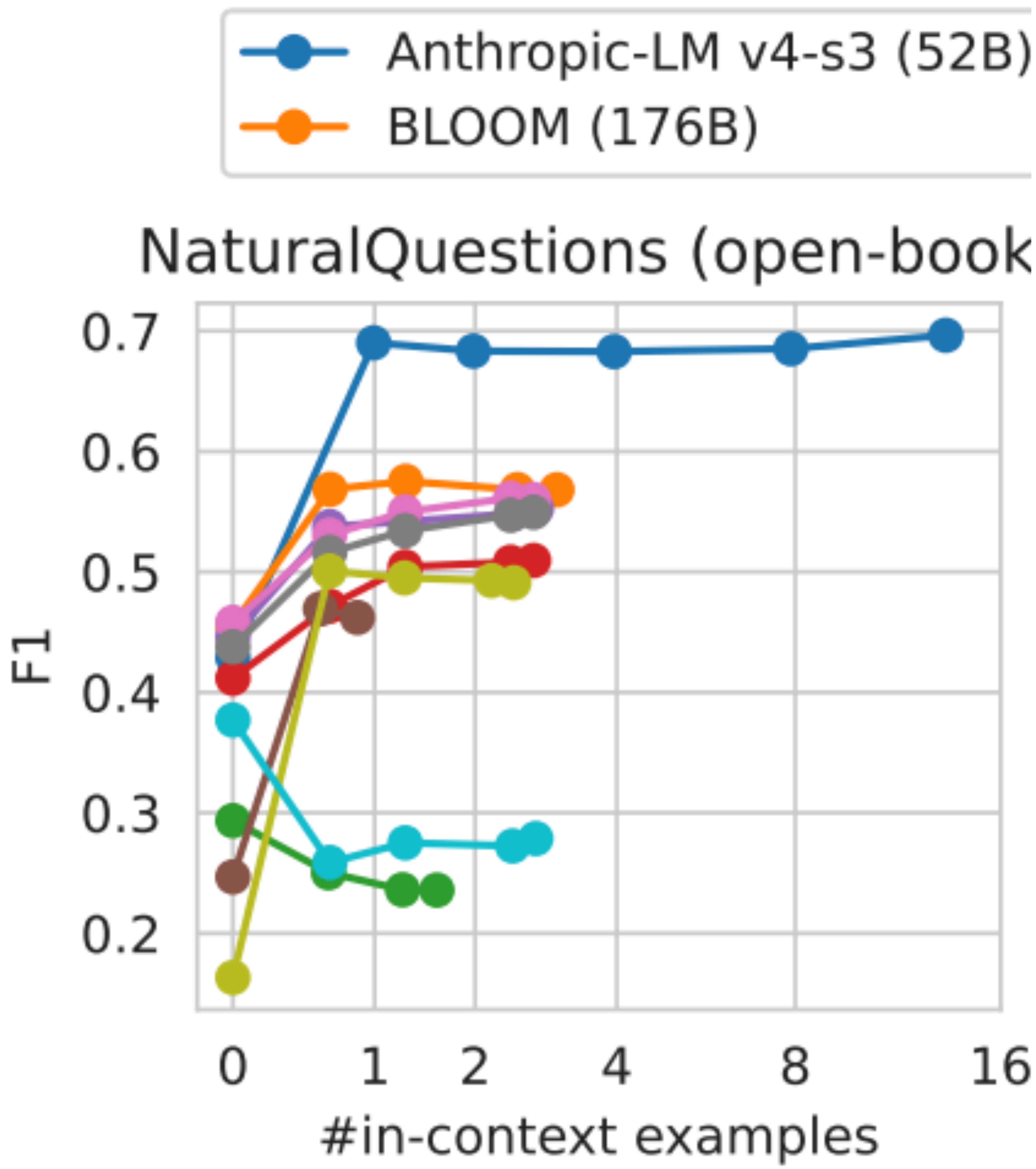


- Solution: "calibrate" the model by normalizing by that probability of null $\boldsymbol{x}_{\text{test}}$

- Leads to higher performance; not necessarily crucial with prompt-tuned models

Zhao et al. (2021)

Results: HELM

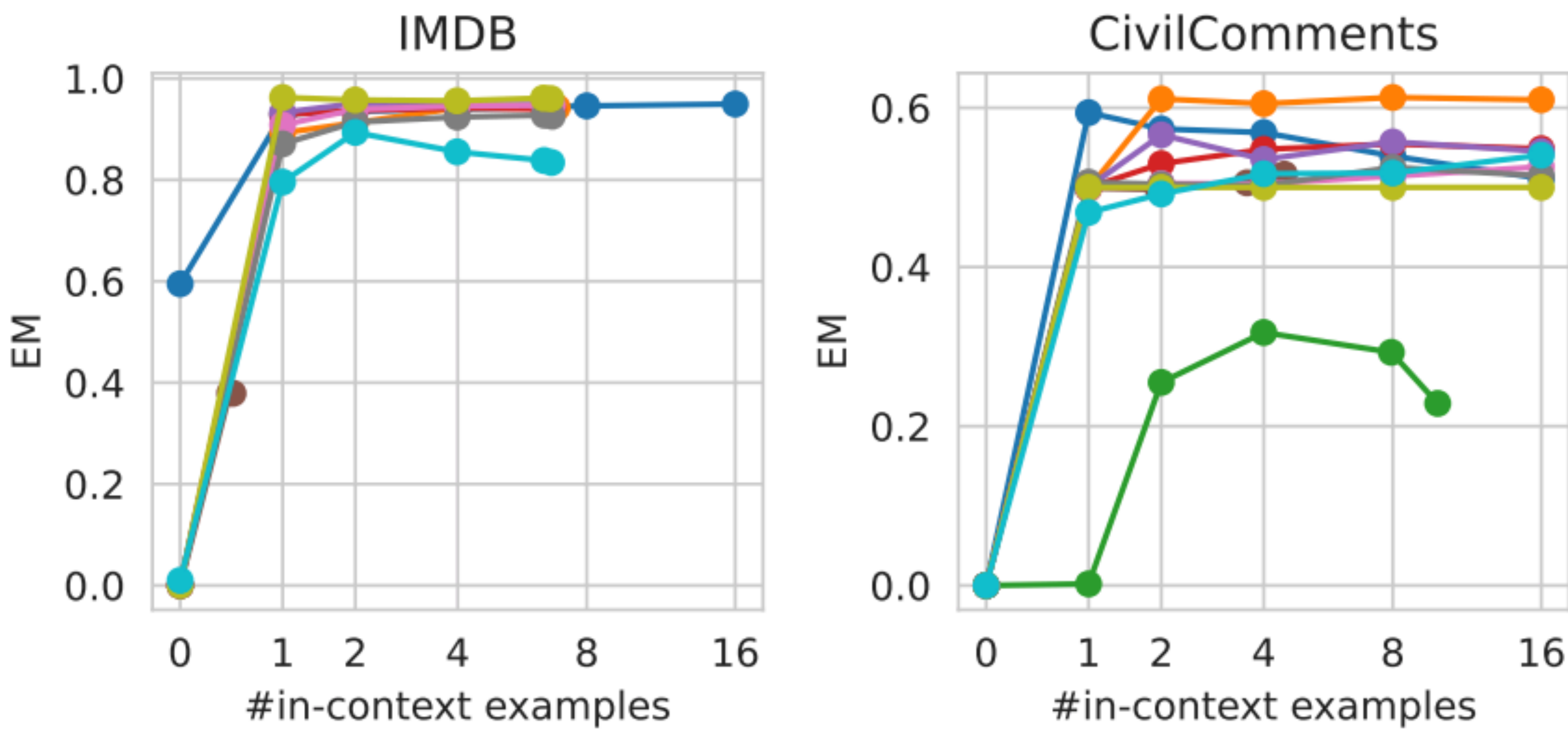


Each line is a different LM

- **More in-context examples generally leads to better performance**

Percy Liang et al. (2022)

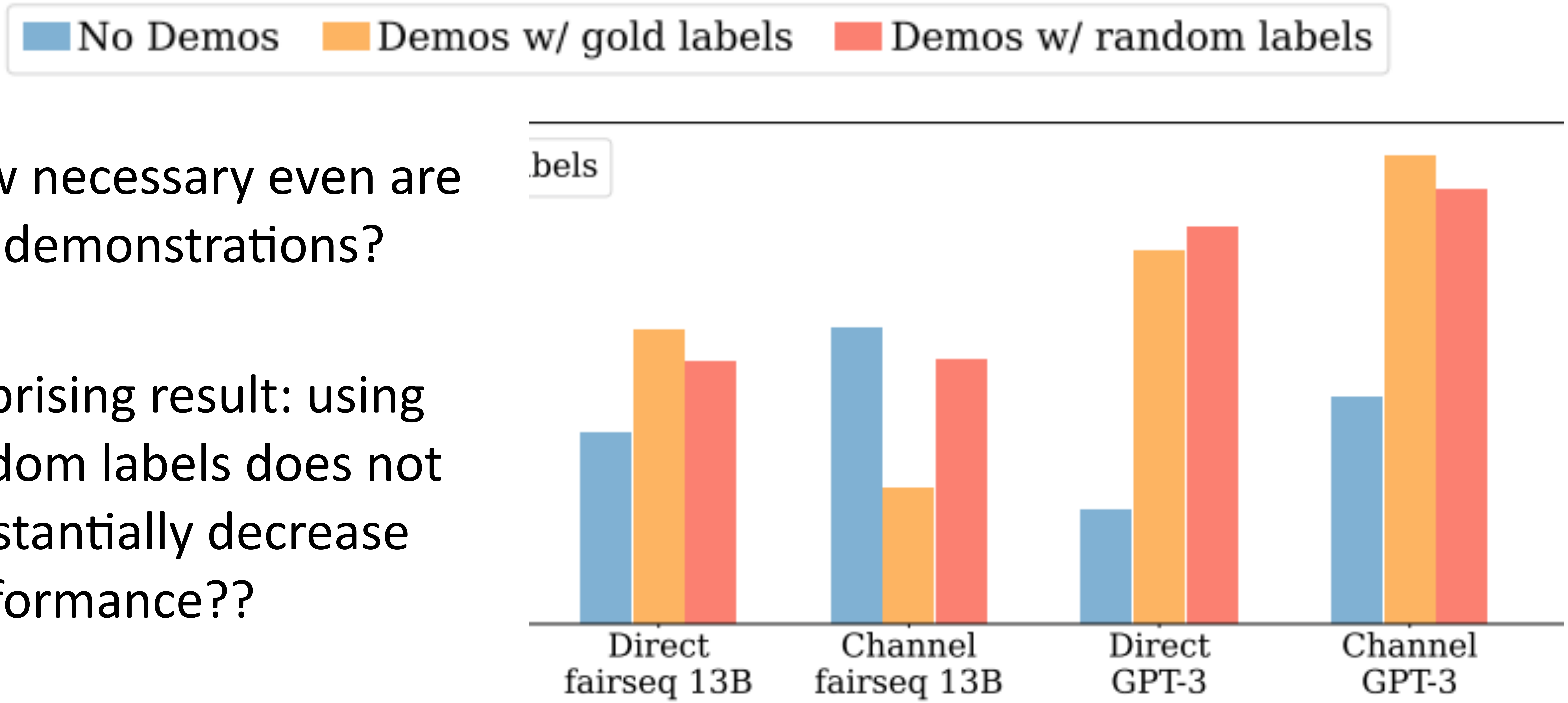
Results: HELM



Percy Liang et al. (2022)

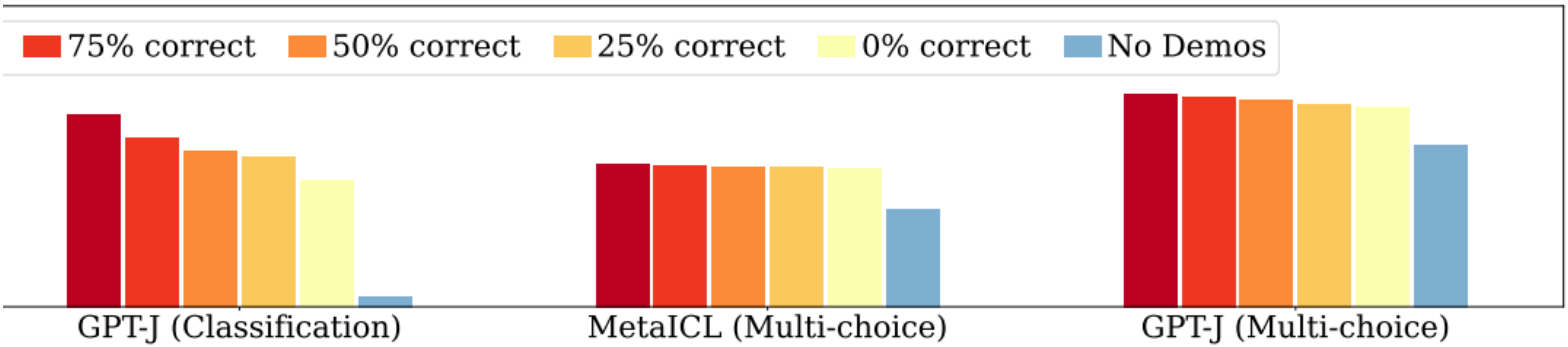
Rethinking Demonstrations

- How necessary even are the demonstrations?
- Surprising result: using random labels does not substantially decrease performance??



Min et al. (2022)

Rethinking Demonstrations



- Having even mislabeled demonstrations is much better than having no demonstrations, indicating that the form of the demonstrations is partially responsible for in-context learning

Min et al. (2022)